TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky a mezioborových inženýrských studií

Studijní program: M 2612 – Elektrotechnika a informatika
Studijní obor: 3906T001 – Mechatronika

Slepá separace řeči ze stereofonního záznamu

Blind speech separation from stereorecording signans

Diplomová práce

Autor: **Michal Kuna**
Vedoucí práce: Ing. Zbyněk Koldovský,Ph.D.
Konzultant: Ing. Jan Kolorenč

V Liberci 17. 5. 2007

Prohlášení

Byl(a) jsem seznámen(a) s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 o právu autorském, zejména § 60 (školní dílo).

Beru na vědomí, že TUL má právo na uzavření licenční smlouvy o užití mé diplomové práce a prohlašuji, že **s o u h l a s í m** s případným užitím mé diplomové práce (prodej, zapůjčení apod.).

Jsem si vědom(a) toho, že užít své diplomové práce či poskytnout licenci k jejímu využití mohu jen se souhlasem TUL, která má právo ode mne požadovat přiměřený příspěvek na úhradu nákladů, vynaložených univerzitou na vytvoření díla (až do jejich skutečné výše).

Diplomovou práci jsem vypracoval(a) samostatně s použitím uvedené literatury a na základě konzultací s vedoucím diplomové práce a konzultantem.

Datum

Podpis

Poděkování

Na tomto místě chci poděkovat především vedoucímu diplomové práce Ing. Zbyňkovi Koldovskému,Ph.D. za ochotu a odbornou pomoc, kterou mi ochotně poskytoval v průběhu řešení práce. Mé poděkování patří také Ing. Jindřichovi Žďánskému,Ph.D. za testování vyčištěných signálů na automatickém rozpoznávání řeči. Tím mi pomohl s realizací experimentu, který ověřil účinnost navržené metody. Zároveň děkuji rodině za materiální i psychickou podporu po celou dobu studia.

Abstrakt

Diplomová práce se zabývá návrhem efektivního algoritmu pro slepou separaci signálu řeči ze stereofonního záznamu v časově-frekvenčním pásmu s adaptivní volbou prahového parametru τ. Řeč je zarušena stereofonní hudbou v pozadí.

Anotace

Úkolem práce je navrhnout efektivní algoritmus pro slepou separaci signálu (BSS – blind signal separation) v časově-frekvenčním pásmu s adaptivní volbou prahového parametru τ. Konkrétně je úkolem práce návrh algoritmu pro separaci řeči ze stereofonního záznamu zarušeného hudbou v pozadí. Metoda je založena na základě odhadu vzájemné informace vstupních signálů a na předpokladu, že vstupní signály, respektive levá a pravá strana stereofonního signálu, jsou alespoň částečně nezávislé. To znamená, že mají minimální vzájemnou informaci. Čím větší je vzájemná nezávislost vstupních signálů, tím je algoritmus účinnější. Původní signály jsou neznámé a jsou navzájem určitým způsobem smíšeny. Obecně se mixují pomocí takzvané mixovací matice, která je v čase proměnná. V našem případě se však jedná o prostý součet ovlivněný pouze intenzitou jednotlivých signálů. Signál řeči je přičten k oběma kanálům hudby stejně. Ze vzájemné informace vstupních signálů určíme parametry separace tak, aby byla co nejefektivnější.

Na základě simulací navrhneme optimální volbu dalších parametrů metody. Separovaný signál totiž není možno úplně zrekonstruovat, ale optimálním nastavením parametrů se můžeme úplné rekonstrukci velmi přiblížit.

Abstract

This graduation theses put mind on suggestion of effective algorithm for blind speech signal separation from stereophonic recording in time-frequency band with adaptive choise of threshold parameter τ. Speech is mixing with stereophonic music in background.

Anotation

Task of this work is to propose effective algorithm for Blind Signal Separation (BBS) in time-frequency band with adaptive choice of threshold parameter τ. In concrete terms is task of the work algorithm desing for the purpose is speech separation from stereophonic recording mixing by music in background. The method is based on estimation of reciprocal infomration of entrance signals and on assumption of partitally independent entrance signals (left and right side of music signal). It means that they have minimal reciprocal information. More the bertter the reciprocal infomration of entrance signals the algorithm is more efectively. The original signals are unknown and in a way mixed. In general terms are signals mixed with so called time variable mixing matrix. In our case is acting as simply summation affected only by intensity of single signals. The speech signal si added to both sides of music signals in the same way. From reciprocal information entrance signals we assess separation parameters so that it would be to most effective.

On the basis of simulations we suggest optimal choice of next method parameters. Obtained signal is impossible to completely reconstruct, but we can come near to complete reconstruction by optimal parameters settings.

Obsah

Seznam obrázků

Seznam grafů

1. Úvod

Řešená úloha se zabývá problémem souvisejícím s dnes moderním digitálním rozpoznáváním řeči a jejím počítačovým zpracováním. Toto je dále využíváno pro řadu aplikací jako je hlasové ovládání počítačů, automatický přepis hlasových záznamů a podobně. Pro všechny tyto aplikace je nutné hlasový záznam oddělit od všech ostatních rušivých signálů snižujících kvalitu záznamu.

Tato práce se zabývá návrhem metody pro řešení separace řeči ze stereofonního záznamu, kde je řeč rušena stereofonní hudbou různé intenzity v pozadí. Mým úkolem bylo navrhnout efektivnější a jednodušší algoritmus, než je současná metoda.

Úkolem je navrhnout efektivní algoritmus pro slepou separaci signálu (BSS – Blind Signal Separation) v časově-frekvenčním pásmu s adaptivní volbou prahového parametru τ. Nejprve jsem se musel seznámit se stávající metodou separace a poté navrhnout způsob adaptivní volby parametrů a optimální algoritmus separace tak, abych dosáhl co nejlepších výsledků.

V první části práce je definována vzájemná informace dvou nezávislých veličin, kterou budeme dále využívat pro určení vzájemné závislosti levé a pravé strany stereofonního signálu. A k tomu jsou samozřejmě definovány související pojmy z teorie pravděpodobnosti.

Původní signály řeči a hudby nejsou předem známy a jsou vzájemně míšeny. Mixování je v tomto případě provedeno pouze jednoduchým součtem ovlivněným jen intenzitou jednotlivých signálů. Jejich poměr též neznáme a může se výrazně pro jednotlivé signály lišit. Současně mohou být signály hudby i řeči velmi rozmanité co do frekvenčního spektra i časového rozložení. Z toho vyplývá, že řešení úlohy bude značně nejednoznačné. To znamená, že optimální hodnoty jednotlivých parametrů se budou pro jednotlivé konkrétní signály výrazně lišit, a bude proto obtížné vhodně zvolit jednotlivé parametry tak, aby řešení vyhovovalo pro co nejširší spektrum různých signálů.

Na základě simulací navrhneme optimální volbu dalších parametrů metody tak, abychom se co nejvíce přiblížili původnímu signálu řeči. Musíme najít takové nastavení, abychom maximálně potlačili hudbu, ale zachovali řeč bez výrazného

zkreslení. Signál není možné zrekonstruovat ideálně, ale optimálním nastavením parametrů se můžeme ideální rekonstrukci ve smyslu některého kritéria velmi přiblížit.

Na závěr práce provedeme experimentální ověření správnosti navržené metody separace pomocí automatického přepisu známých zvukových záznamů. Naší snahou v tomto experimentu bude co nejvíce se přiblížit původnímu záznamu řeči v procentuální úspěšnosti přepisu. To znamená, aby došlo oproti vstupnímu zarušenému signálu k co největšímu zlepšení. Výsledky práce budou dále využity pro další aplikace zpracování řeči v laboratoři zpracování řeči na fakultě mechatroniky Technické univerzity v Liberci.

2. Vzájemná informace

2.1. Úvod

Na vzájemné informaci levého a pravého kanálu vstupního signálu výrazně závisí prahový parametr τ. Jak se později ukáže, τ je důležitý parametr výrazně ovlivňující kvalitu separace. Podrobněji se jím budeme zabývat v kapitole 3, která se zabývá vlastní separací signálů.

Vzájemná informace je určena značně komplikovaným integrálem, který nelze řešit analyticky, protože v praxi neznáme hustoty pravděpodobností vstupních signálů. Přímý odhad vzájemné informace tedy provádíme pomocí numerického řešení tohoto integrálu.

2.2. Základní pojmy z teorie pravděpodobnosti

Mějme prostor náhodných jevů Ω. Potom náhodná veličina X: Ω→R je takové zobrazení, že množina $\{\omega|X(\omega) \leq x\}$ je pro všechna x borelovsky měřitelná. Můžeme tedy definovat její distribuční funkce F: R→R kde $F_x(x) = P(X \leq x)$ pro všechna $x \in R$. Hustotu pravděpodobnosti f_X potom definujeme jako funkci, která splňuje $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$, pokud taková funkce existuje.

Pro n náhodných veličin $X_1,\ldots, X_n$ pak definujeme hustoty pravděpodobnosti $f_{X1},\ldots,f_{Xn}$. Pokud jsou $X_1,\ldots, X_n$ nezávislé, pak platí, že jejich sdružená hustota pravděpodobnosti $f_{X1,\ldots,Xn}$ je:

$$f_{X1,\ldots,Xn}(X_1,\ldots,X_n) = \prod_{i=1}^{n} f_{Xi}(X_i) \tag{1}$$

Střední hodnota E a rozptyl D náhodné veličiny X jsou definovány:

$$EX = \int_R x f_X(x)dx \tag{2}$$

$$DX = E(X - EX)^2 \tag{3}$$

Kovariance dvou náhodných veličin X a Y je definována jako:

$$Cov(X,Y) = E\left[(X - EX)(Y - EY)\right] = E(XY) - EXEY \tag{4}$$

Nutnou podmínkou, aby X a Y byly statisticky nezávislé, je: Cov(X, Y) = 0.

2.3. Definice vzájemné informace

Vzájemná informace náhodných veličin $X_1,\ldots,X_n$ je definována jako Kullback-Leiblerova divergence mezi sdruženou hustotou pravděpodobnosti $fx_1,\ldots,x_n$ a součinem hustot pravděpodobností jednotlivých proměnných $\prod_{i=1}^{n} f_{Xi}$:

$$I(X_1,...,X_n) = \int_{R^n} f(x_1,...,x_n) \ln \frac{fx_1,...,x_n(x_1,...,x_n)}{\prod_{i=1}^{n} fx_i(x_i)} dx_1...dx_n \tag{5}$$

Pro analytický výpočet vzájemné informace tedy potřebujeme znát marginální a sdružené hustoty pravděpodobností vstupních signálů. V našem případě, kdy vstupními proměnnými jsou signály levého a pravého kanálu stereofonního záznamu, tyto pravděpodobnosti neznáme. Proto nejsme schopni analyticky vzájemnou informaci vstupních signálů určit, a musíme tedy přistoupit k jejímu odhadu.

2.4. Přímý odhad vzájemné informace

Výpočet spočívá v numerickém odhadu integrálu (5):

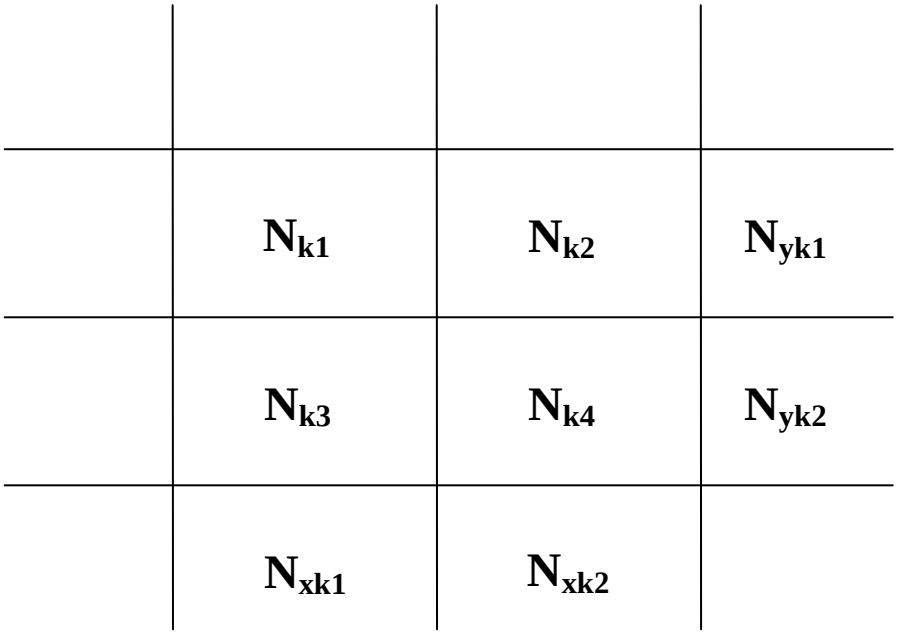
$$I(X_1,...,X_n) = \int_{R^n} f(x_1,...,x_n) \ln \frac{fx_1,...,x_n(x_1,...,x_n)}{\prod_{i=1}^{n} fx_i(x_i)} dx_1...dx_n \tag{5}$$

Odhad integrálu, jak byl navržen v [4], vychází z pozorování náhodných veličin $X_1,\ldots, X_n$. Algoritmus výpočtu bude popsán pro dimenzi 2 (n = 2), neboť pro tento případ je výpočet nejnázornější, a pro vyšší dimenze je analogický. V našem případě máme pouze dvě vstupní proměnné, jimiž jsou vstupní signály.

Postup lze rozdělit do tří bodů:

1. Zobrazení všech pozorování v jednom obdélníku
2. Každý obdélník rozdělíme horizontálně a vertikálně tak, aby vzniklé obdélníky byly vzhledem k marginálním (okrajovým) distribucím stejně pravděpodobné. To znamená, že $N_{xk1} = N_{xk2}$ a $N_{yk1} = N_{yk2}$.
3. Pokud hodnota $fx_1, x_2(x, y) / [fx_1(x)\, fx_2(y)]$ není ve vzniklých obdélnících stejná, dělíme každý nový obdélník stejně jako v druhém bodě.

N je počet pozorování náhodné veličiny $X = (X_1, \ldots, X_n)$. N_{xk} a N_{yk} jsou tedy počty pozorování ve směru x a y. Dělení obdélníku můžeme zobrazit na následujícím obrázku (Obr.1)



Obr.1: Rozdělení k-tého obdélníku

Za předpokladu, že jsme N_{xk} a N_{yk} přesně rozdělili, platí $N_{yk1} = N_{yk2} = N_{yk}/2$ a $N_{xk1} = N_{xk2} = N_{xk}/2$. Požadavek je, aby:

$$\frac{f_{X,Y}(x,y)}{f_X(x)\, f_Y(y)} = \frac{N_{kj}/N}{(N_{xkj}/N)(N_{ykj}/N)} = N_{kj}\left(\frac{4N}{N_{xk}N_{yk}}\right) \tag{6}$$

byla pro každé $j \in \{1,...,4\}$ stejná. Jako test hypotézy použijeme kvadratické kritérium.

$$\frac{4}{N_k}\sum_{j=1}^{4}\frac{N_k}{N}\log\frac{N_k/N}{N_{xk}/N\cdot N_{yk}/N}\leq 7{,}9 \tag{7}$$

Prostor měření je tedy rozdělen do obdélníků obsahujících méně než čtyři měření, nebo do obdélníků s lokální nezávislostí, kde je přibližně splněna podmínka (6). Konečný vztah pro odhad vzájemné informace tedy bude vypadat takto:

$$I(X,Y)\approx\sum_{k=1}^{K}\frac{N_k}{N}\log\frac{N_k/N}{N_{xk}/N\cdot N_{yk}/N} \tag{8}$$

kde k je počet dělení obdélníků, N je počet pozorování náhodné veličiny $X = (X_1, \ldots, X_n)$. N_k je celkový počet pozorování v k-tém obdélníku. N_{xk} a N_{yk} jsou počty pozorování ve směru x a y.

3. Separace řeči ze stereofonního záznamu

Slepá separace signálů se obecně zabývá problémem, jak z nějakého vstupního signálu, který vznikl smícháním obecně libovolného počtu různých signálů, získat zpět tyto původní signály. V obecném případě předem neznáme žádný ze vstupních signálů ani způsob, jakým jsou mixovány. Proto hovoříme o slepé separaci. Každý signál můžeme chápat jako náhodný proces.

V našem případě se jedná o separaci řeči ze stereofonního záznamu. Jde o speciální aplikaci slepé separace signálů, která se snaží o získání řeči ze záznamu s minimálním zkreslením a o maximální potlačení hudby.

Stereofonní signál se skládá z takzvaného levého a pravého kanálu. Levý a pravý kanál hudby jsou částečně odlišné záznamy a mezi nimi uprostřed je řeč. Řeč je k oběma signálům přičtena se stejnou intenzitou.

3.1. Základní předpoklady

Dva signály, levou a pravou stranu záznamu, považujeme za vstupní proměnné x(t). Popsat je můžeme následujícími vztahy (9):

$$\text{L: } x_1(t) = s(t) + y_1(t) \tag{9}$$

$$\text{P: } x_2(t) = s(t) + y_2(t)$$

kde vektor $x(t) = [x_1(t), x_2(t)]^T$ je vektor získaných mixovaných signálů, $y(t) = [y_1(t), y_2(t)]^T$ je vektor původních signálů levé a pravé strany hudby. s(t) je původní signál řeči, který je přičten k oběma hudebním kanálům. Snažíme se co nejvěrněji zrekonstruovat signál s(t) a zároveň potlačit signál y(t). Index vzorku t = 1,…,N, kde N je počet samplů jednotlivých signálů.

Předpokládáme, že s(t) je nezávislý na $y_1(t)$ a $y_2(t)$. Ani jeden z těchto třech signálů neznáme v čisté podobě, proto tento problém spadá do slepé separace. Signály $y_1(t)$ a $y_2(t)$ mohou být vzájemně částečně závislé, ale nesmějí být identické. V ideálním případě jsou $y_1(t)$ a $y_2(t)$ nezávislé. Čím je výraznější rozdíl mezi $y_1(t)$ a $y_2(t)$, tím jsme schopni je lépe potlačit a tím věrněji obnovit signál s(t). Můžeme předpokládat, že

všechny signály mají nulovou střední hodnotu (10), což nám zjednoduší pozdější výpočty:

$$E[s(t)] = E[y_1(t)] = E[y_2(t)] = 0 \tag{10}$$

3.2. Postup při řešení

Základním prvkem řešení je určení signálů u(t) a v(t). Tyto signály vytvoříme součtem, respektive rozdílem $x_1(t)$ a $x_2(t)$. Tím zajistíme, že je v signálu u(t) maximální poměr energie řeči a hudby a v signálu v(t) je signál řeči s(t) úplně potlačen.

$$\begin{aligned} u(t) &= \frac{1}{2}\left(x_1(t) + x_2(t)\right) = s(t) + \frac{1}{2}y_1(t) + \frac{1}{2}y_2(t) \\ v(t) &= x_1(t) - x_2(t) = y_1(t) - y_2(t) \end{aligned} \tag{11, 12}$$

Protože signály s(t) a y(t) jsou vzájemně nezávislé, je zřejmé, že tímto způsobem vytvořené signály u(t) a v(t) jsou vzájemně maximálně nezávislé. Míra jejich nezávislosti závisí na poměru energií původních signálů s(t) a y(t). V případě, že je ve vstupním signálu x(t) signál s(t) výrazně zastoupen, rozdíl mezi signály u(t) a v(t) je výrazný a signály mají velkou míru nezávislosti. Jestliže však je ve vstupním signálu x(t) signál s(t) obsažen jen minimálně nebo vůbec, signály u(t) a v(t) jsou značně závislé, protože i původní signály $y_1(t)$ a $y_2(t)$ jsou často závislé.

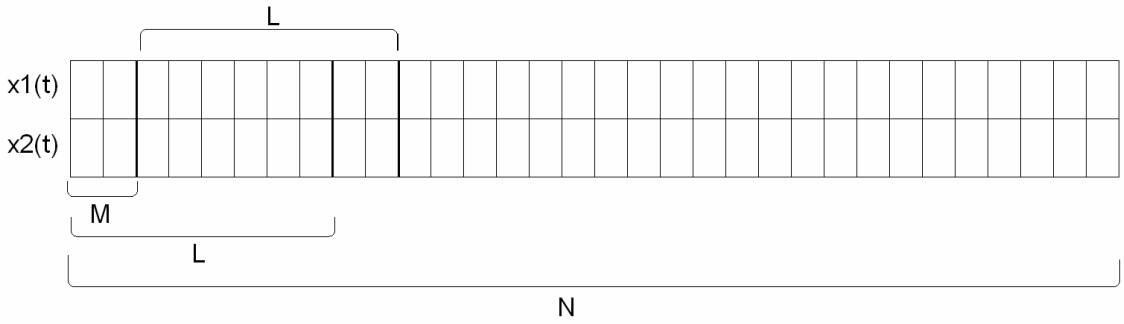
Pro získání signálu s(t) se používá maskování, ale velmi záleží na volbě parametrů masky. Maskování lze provádět pomocí binární masky, ale jako výhodnější se jeví použití spojité masky. Maska nuluje kmitočty v časofrekvenčním pásmu, které chceme odstranit. Výběr kmitočtů, které chceme potlačit, se provádí na základě porovnání signálů u(t) a v(t). Princip maskování popíšeme v kapitole 3.3. Pomocí Short-time Fourierovy transformace (STFT) získáváme obrazy signálů v časofrekvenční oblasti. Kvalita maskování je závislá především na volbě prahového parametru τ. Jeho nevhodná volba může degradovat celou metodu a zkreslit signál s(t). Naopak vhodná volba ovlivní výsledek separace velmi kladně. V případě, že je parametr τ příliš nízký, dojde k minimálnímu potlačení hudby. Naopak v případě, že je τ příliš vysoké, současně s hudbou potlačíme i část řeči, čímž záznam znehodnotíme.

3.2.1. Short-time Fourierova transformace

L-bodovou Short-time Fourierovu transformaci pro signál x(t), kde t = 1,…,N definujeme takto:

$$STFT[x(t)](\omega_k, l) = x(\omega_k, l) = \sum_{i=lM+1}^{lM+L} x(i)\omega(i)e^{-\frac{2\pi j}{L}(i-lM-1)(\omega_k-1)} \tag{13}$$

kde L je délka okénka, M je délka nepřekrývajících se částí sousedních okének, l = 0,…,(p-1)r ω_k = 1,…,L, r = L/M a p = N/L. Předpokládáme, že konstanty p a r jsou celá čísla. Z praxe je zřejmé, že na konec záznamu můžeme přidat libovolně dlouhý nulový signál, který je možné po provedení separace opět odstranit. Potom bude délka signálu N taková, že umožňuje splnění tohoto předpokladu. Později se ukáže, že nejvýhodnější je použití velikosti okénka L = 1024 a velikosti nepřekrývajících se částí M = 128. Okénkovací funkci ω volíme ω = 1, protože složitější okénko by nepřineslo větší přesnost, ale pouze by zvýšilo výpočtovou náročnost.



Obr.2: Překryv sousedních okének

3.2.2. Časofrekvenční maskování

Nechť M(ω_k, l) je kladná reálná funkce reprezentující masku v časofrekvenční oblasti, a nechť s(ω_k, l), y_1(ω_k, l), y_2(ω_k, l), u(ω_k, l) a v(ω_k, l) jsou obrazy Short-time Fourierovy transformace signálů s(t), y_1(t), y_2(t), u(t) a v(t) v časofrekvenční oblasti. Maskované signály jsou pak definovány jako inverzní Short-time Fourierova

transformace fourierova obrazu signálu v časofrekvenční oblasti vynásobeného maskovací funkcí M(ω_k, l):

$$\tilde{s}^M(t) = ISTFT\left[M(\omega_k, l)s(\omega_k, l)\right](t) \quad (14)$$

$$\tilde{y}_1^M(t) = ISTFT\left[M(\omega_k, l)y_1(\omega_k, l)\right](t) \quad (15)$$

$$\tilde{y}_2^M(t) = ISTFT\left[M(\omega_k, l)y_2(\omega_k, l)\right](t) \quad (16)$$

Hledaný signál $\hat{s}^M(t)$ vypočteme jako inverzní Short-time Fourierovu transformaci signálu u(ω_k, l) vynásobeného maskou M(ω_k, l):

$$\hat{s}^M(t) = ISTFT\left[M(\omega_k, l)u(\omega_k, l)\right](t) \quad (17)$$

3.3. Maskování signálu

Postup potlačení hudby pomocí určité masky, jak byl navržen v [6], vychází z předpokladu splnění podmínky takzvané W-disjoint orthogonality původních signálů. To znamená, že v každém bodě (ω_k, l) časofrekvenční oblasti je vždy pouze jeden z původních signálů s(ω_k, l), nebo y(ω_k, l) nenulový. y(ω_k, l) je vektor reprezentující stereofonní signál hudby v časofrekvenční oblasti:

$$y(\omega_k, l) = [y_1(\omega_k, l), y_2(\omega_k, l)]^T \quad (18)$$

Můžeme použít dva typy masky: binární a spojitou. Obě masky jsou v limitních situacích shodné. Spojitá maska oproti binární nemá skokovou změnu mezi krajními stavy, ale přechod je pozvolný.

Binární maska:

$$M(\omega_k, l) = \begin{cases} 1 & |u(\omega_k, l)| > \tau |v(\omega_k, l)| \\ 0 & jindy \end{cases} \quad (19)$$

Binární maska porovnává signály u(t) a v(t) s ohledem na prahový parametr τ. Vztah (19) popisující binární masku vyplývá z předpokladu W-disjoint orthogonality vstupních signálů a skutečnosti, že signál u(t) v sobě obsahuje maximální energii signálu s(t), zatímco v signálu v(t) signál s(t) není obsažen vůbec.

V případě, že je předpoklad W-disjoint orthogonality splněn bezezbytku, mohou nastat pouze dvě situace.

1) $y(t) \neq 0$, $s(t) = 0$: Signál u(t) obsahuje pouze součet složek $y_1(t)$ a $y_2(t)$:

$$u(t) = \frac{1}{2} y_1(t) + \frac{1}{2} y_2(t)$$

Signály $y_1(t)$ a $y_2(t)$ bývají často velmi podobné, proto bude $u(\omega_k, l)$ podobný $v(\omega_k, l)$ a binární maska tento bod (ω_k, l) v časofrekvenční oblasti potlačí.

2) $y(t) = 0$, $s(t) \neq 0$: Potom naopak signál u(t) obsahuje pouze s(t) a signál v(t) je nulový. Z toho plyne, že $|u(\omega_k, l)|$ bude vždy větší než $|v(\omega_k, l)|$. Tento bod ponecháme nezměněný.

Z definice signálů u(t) a v(t) je zřejmé, že signál $u(\omega_k, l)$ bude vždy silnější, než $v(\omega_k, l)$, proto při jejich srovnání uvažujeme vždy jen $|u(\omega_k, l)| > |v(\omega_k, l)|$. Prahový parametr τ určuje práh, od kterého už rozdíl mezi signály $u(\omega_k, l)$ a $v(\omega_k, l)$ považujeme za natolik významný, abychom daný bod v časofrekvenční oblasti zachovali.

V případě, že se τ v limitě blíží nekonečnu, pak potlačíme veškerý hudební signál, ale zároveň i řeč. V opačném případě, kdyby τ bylo nulové, naopak nedojde k žádnému potlačení a signál u(t) zůstane nezměněn. Adaptivní volba prahového parametru τ se zabývá právě jeho vhodným určením tak, abychom potlačili maximum hudby a přitom zachovali signál řeči s(t). Adaptivní volba parametru τ bude popsána v kapitole 3.5.

Binární maska je nejjednodušší a v případě splnění předpokladu W-disjoint orthogonality funguje nejlépe, ovšem má svá omezení. Hlavní nevýhodou je, že předpoklad W-disjoint orthogonality není v reálných aplikacích beze zbytku splněn. To znamená, že se mohou objevit body v časofrekvenční oblasti pro které bude platit, že oba z původních signálů $s(\omega_k, l)$ a $y(\omega_k, l)$ budou nenulové.

Dalším problémem je ostrý přechod mezi signálem, který je potlačen, a signálem, který zachováme, protože maska funguje tak, že daný bod signálu v časofrekvenční oblasti buď nulujeme, nebo zachováme nezměněný. Proto se v reálných aplikacích používá sice složitější, ale vhodnější spojitá maska. Její výhodou je, že řeší nedostatky binární masky.

Spojitá maska:

$$M(\omega_k, l) = \frac{|u(\omega_k, l)|^2}{|u(\omega_k, l)|^2 + \tau |v(\omega_k, l)|^2} \tag{20}$$

Tento tvar masky byl navržen v [1]

Hledáme-li minimum funkce:

$$|s(\omega_k, l) - M^i(\omega_k, l) u(\omega_k, l)|^2 \tag{21}$$

Jednoduchým výpočtem pak získáme vztah pro ideální masku $M^i(\omega_k, l)$:

$$M^i(\omega_k, l) = \frac{|s(\omega_k, l)|^2 + \Re\left(\overline{s(\omega_k, l)} \bar{y}(\omega_k, l)\right)}{|u(\omega_k, l)|^2} \tag{22}$$

kde $\bar{y}(\omega_k, l) = \frac{y_1(\omega_k, l) + y_2(\omega_k, l)}{2}$ a $\Re(z)$ reprezentuje reálnou část komplexního čísla z. Z výsledného vztahu pro ideální masku (22) můžeme vidět, proč byl navržen vztah (20) popisující spojitou masku.

Prahový parametr τ ovlivňuje separaci pomocí spojité masky obdobným způsobem jako u binární masky. V případě, že se τ bude v limitě blížit nekonečnu, bude se jmenovatel vztahu (20) též blížit nekonečnu. Z toho plyne, že celý vztah bude roven nule. Dojde tedy k totálnímu potlačení vstupního signálu. Naopak v případě, že τ bude nulové, vztah (20) se zjednoduší na:

$$M(\omega_k, l) = \frac{|u(\omega_k, l|^2}{|u(\omega_k, l|^2} = 1 \tag{23}$$

To znamená, že zůstane zachován veškerý vstupní signál, stejně jako při použití binární masky s $\tau = 0$.

Ze vztahů popisujících obě masky je zřejmé, že na kvalitu výsledného signálu bude mít zásadní vliv prahový parametr τ. Proto se jako jeden z nejdůležitějších problémů ukazuje právě volba τ.

3.4. Kritéria kvality separace

Kvalitu získaného záznamu je možno posuzovat podle několika částečných kritérií. Jedná se o DSR (Distortion to Signal Ratio) udávající zkreslení získaného signálu $\hat{s}^M(t)$ vůči původnímu s(t) a ISR (Interference to Signal Ratio), který udává poměr energie signálu y(t) vůči signálu s(t). V podstatě udává potlačení signálu y(t).

$$DSR^M = \frac{\min_\alpha E\left[s(t) - \alpha\tilde{s}^M(t)\right]^2}{E\left[s(t)\right]^2} \tag{24}$$

$$ISR^M = \frac{\min_\alpha E\left[\hat{s}^M(t) - \alpha\tilde{s}^M(t)\right]^2}{E\left[\tilde{s}^M(t)\right]^2} \tag{25}$$

Tato kritéria udávají však vždy jen jeden náhled na kvalitu separace a to buď zkreslení signálu s(t) nebo potlačení y(t). My však potřebujeme situaci posoudit oběma těmito pohledy a najít nejvýhodnější kompromis mezi nimi.

Proto bude zřejmě nejvýznamnější kritérium IDSR (Interference plus Distortion to Signal Ratio). To udává poměr energie zkreslení signálu $\hat{s}^M(t)$ a potlačení y(t) vůči původnímu signálu s(t)

$$IDSR^M = \frac{\min_\alpha E\left[s(t) - \alpha\hat{s}^M(t)\right]^2}{E\left[s(t)\right]^2} \tag{26}$$

kde E[.] je střední hodnota náhodné veličiny, v našem případě signálu. Tyto vztahy však nejsme schopni analyticky vyřešit.

Protože Short-time Fourierova transformace je lineární transformací, což platí samozřejmě i o inverzní transformaci, platí vztah:

$$\hat{s}^M(t) = \tilde{s}^M(t) + \frac{1}{2}\tilde{y}_1^M(t) + \frac{1}{2}\tilde{y}_2^M(t) \tag{27}$$

Za předpokladu, že signál s(t) je nezávislý na $y_1(t)$ a $y_2(t)$, pak můžeme podmínky přepsat do tvaru, který je již prakticky použitelný:

$$IDSR^M = 1 - \frac{\hat{E}^2\left[s(t)\tilde{s}^M(t)\right]}{\hat{E}\left[s(t)\right]^2 \hat{E}\left[\hat{s}^M(t)\right]^2} \tag{28}$$

$$DSR^M = 1 - \frac{\hat{E}^2\left[s(t)\tilde{s}^M(t)\right]}{\hat{E}\left[s(t)\right]^2 \hat{E}\left[\tilde{s}^M(t)\right]^2} \tag{29}$$

$$ISR^M = \frac{\hat{E}\left[\tilde{y}_1^M(t) + \tilde{y}_2^M(t)\right]^2}{4 \cdot \hat{E}\left[\tilde{s}^M(t)\right]^2} \tag{30}$$

kde $\hat{E}[\cdot]$ je výběrová střední hodnota signálu (průměr). Pomocí těchto vztahů jsme již schopni relativně snadno určit hodnoty kritérií. Největší význam pro nás má IDSR, protože ho využijeme dále jako kritérium kvality separace při hledání vhodné hodnoty optimálního parametru τ.

Žádné z těchto kritérií kvality separace nejsme v reálné aplikaci schopni použít, protože neznáme původní signál s(t), který je pro jejich výpočet zapotřebí. IDSR využijeme pouze v laboratorních podmínkách pro experiment, jehož výsledkem bude vhodná volba parametrů navržené metody separace. Experiment provedeme na uměle vytvořených signálech. Ty vyrobíme zarušením známých záznamů řeči, což nám zajistí znalost signálu s(t). Během experimentu budeme z původních signálů počítat IDSR. To pak využijeme pro vyhodnocení experimentu a volbu nejvhodnějších parametrů metody.

3.5. Adaptivní volba prahového parametru τ

Adaptivní volba prahového parametru τ je způsob určení co nejoptimálnější hodnoty parametru τ. Jeho volba je totiž značně složitá, protože pro každý konkrétní signál se může jeho optimální hodnota τ^{opt} značně lišit. Při experimentu jsem zjistil, že se může pohybovat od nepatrných hodnot blížících se nule, až po hodnoty vyšší než šedesát, výjimečně i více než sto. Takto vysoký rozptyl nám působí značné problémy, chceme-li určit parametr τ nějakým způsobem vhodným pro maximální počet signálů. Navíc není možné τ^{opt} předem určit ze vstupního signálu, aniž bychom měli informace alespoň o jednom z původních signálů s(t), $y_1(t)$ nebo $y_2(t)$.

V další kapitole, která popisuje návrh parametrů separace, bude popsán experiment, jenž ukáže, že hodnota prahového parametru τ významně souvisí se

vzájemnou informací vstupních signálů $x_1(t)$ a $x_2(t)$. Jak bylo již výše popsáno, vzájemná nezávislost signálů $x_1(t)$ a $x_2(t)$ je ovlivněna signálem s(t). Je zřejmé, že čím silnější je signál s(t), tím více klesá vzájemná nezávislost $x_1(t)$ a $x_2(t)$ a roste jejich vzájemná informace. S rostoucím s(t) potřebujeme menší hodnotu prahového parametru τ. Díky tomu je možné určit nějakou funkční závislost τ na vzájemné informaci:

$$\tau = f(I(x)) \tag{31}$$

kde I(x) je velikost vzájemné informace signálů $x_1(t)$ a $x_2(t)$. Experiment také ukáže, že parametr τ je současně závislý i na intenzitě zarušení σ, což vlastně odpovídá poměru intenzit řeči s(t) a hudby $y_1(t)$ a $y_2(t)$.

Experiment spočívá ve vytvoření databáze uměle zarušených signálů. Pro každý z nich pak určíme vzájemnou informaci. Současně hledáme minimální hodnotu funkční závislosti:

$$IDSR = f(\tau) \tag{32}$$

IDSR jsme schopni určit ze znalosti původních signálů, protože databáze byla vytvořena uměle, z důvodu tohoto experimentu. Hodnotu τ, pro kterou je IDSR minimální, považujeme za optimální hodnotu prahového parametru τ^{opt}.

Z výsledku experimentu se nakonec ukáže, že τ^{opt} má pro různé signály značný rozptyl, jak už bylo uvedeno výše. Výsledné hodnoty je zapotřebí vhodným způsobem statisticky zpracovat a nějakým způsobem aproximovat závislost τ na vzájemné informaci (31) tak, abychom dosáhli nejlepších výsledků separace pro maximální možné množství signálů.

Později uvidíme, že τ je na vzájemné informaci nepřímo závislé, a proto bude zřejmě nejrozumnější aproximací pouze jednoduchá lineární závislost τ na převrácené hodnotě vzájemné informace.

4. Volba parametrů metody

Vhodná volba jednotlivých použitých parametrů separace je důležitým krokem, který může výrazně ovlivnit konečný výsledek. Nastavením nevhodných parametrů může dojít ke zkreslení hledaného signálu, nebo dokonce k jeho potlačení, a ztrátě požadované informace.

4.1. Volba masky

Jak bylo již výše zmíněno, maskování lze provádět buď pomocí binární nebo spojité masky. Jednodušší pohled nabízí volba binární masky (19). Ukazuje se ale, že v reálné aplikaci není její použití vhodné, neboť jak již bylo zmíněno v kapitole 3.3., není plně splněn předpoklad W-disjoint orthogonality.

To způsobuje, že s použitím binární masky dochází sice k výraznějšímu potlačení hudby než s použitím spojité masky, ale za cenu příliš vysokého zkreslení s(t). Proto je výhodnější použít spojitou masku (20), která se nejlépe blíží ideální masce a lépe splňuje požadavky reálných aplikací.

4.2. Návrh velikosti okénka a překryvu

Velikost okénka a překryvu, respektive nepřekrývající se části sousedních okének (viz. Obr.2), jsem volil z výsledku provedeného experimentu.

Experiment spočíval v hledání minima IDSR a zápisu odpovídající hodnoty prahového parametru τ, pro různé nastavení velikosti okénka L a nepřekrývající se části sousedních okének M. Experiment jsem provedl pro 1000 vzorků uměle vytvořených signálů, aby byla zajištěna dostatečná objektivita výsledků. Pro každý signál jsem spočítal nejoptimálnější IDSR, které je dosažitelné různou volbou τ.

Tyto výsledky se mohou lišit i velmi výrazně, protože závisí na mnoha dalších faktorech, jako je intenzita řeči, intenzita zarušení, časově - frekvenční spektrum obou

signálů, a podobně. Při každém opakování experimentu jsem vytvořil vzorek signálu tak, že jsem načetl náhodný záznam z databáze nahrávek úryvků řeči ze zpravodajských a publicistických pořadů převážně České televize. Z této nahrávky jsem použil část o délce 2^{15} samplů, s náhodným offsetem od začátku. Databáze mi byla dodána vedoucím diplomové práce a obsahovala 200 různých nahrávek. Délku 2^{15} jsem volil proto, že při vzorkovací frekvenci 16kHz, která byla pro nahrávky použita, odpovídá zhruba dvěma sekundám záznamu. To je postačující pro určení výsledku a současně se nejedná o příliš dlouhý záznam, který by způsobil pouze nárůst výpočtového času, ale nepřinesl by zvýšení přesnosti výsledků. Stejným způsobem jsem vytvořil i náhodný vzorek stereofonní hudby o stejné délce. Databáze hudby obsahovala různé instrumentální záznamy ve formátu .wav, aby byly zpracovatelné Matlabem. Vzorkovací frekvenci záznamů jsem upravil taktéž na 16kHz, aby odpovídala záznamům řeči.

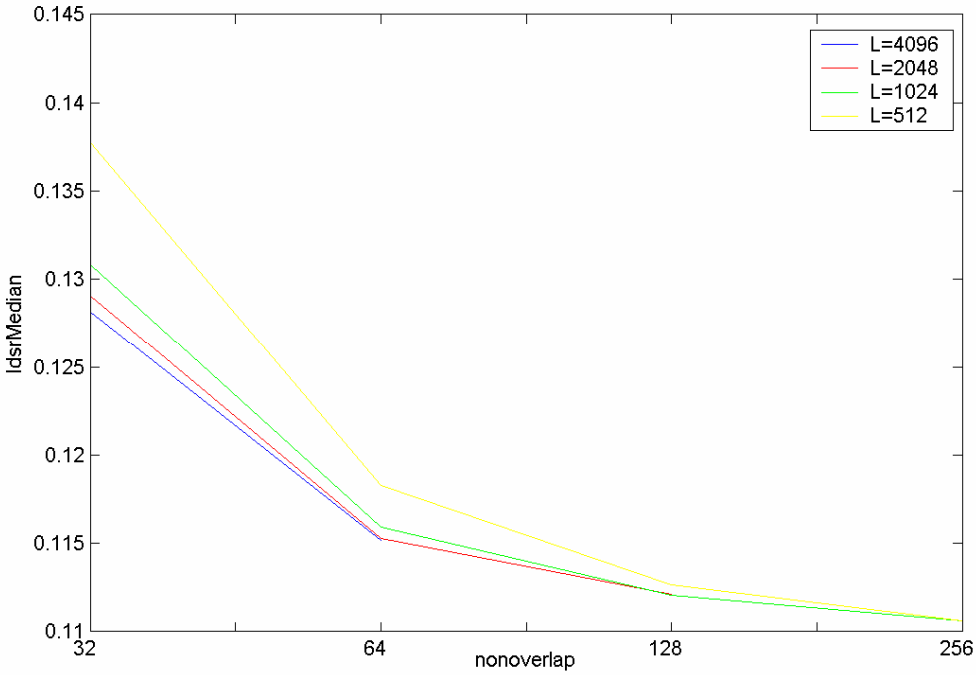
Oba vzorky jsem znormoval, aby měly odpovídající intenzitu a nakonec sečetl v poměru daném parametrem σ, který určoval intenzitu zarušení. Tímto způsobem jsem tedy uměle vytvořil signály, které jsem dále považoval za získané vstupní signály x(t):

$$\begin{aligned} L &: x_1(t) = s(t) + \sigma \cdot y_1(t) \\ P &: x_2(t) = s(t) + \sigma \cdot y_2(t) \end{aligned} \tag{33}$$

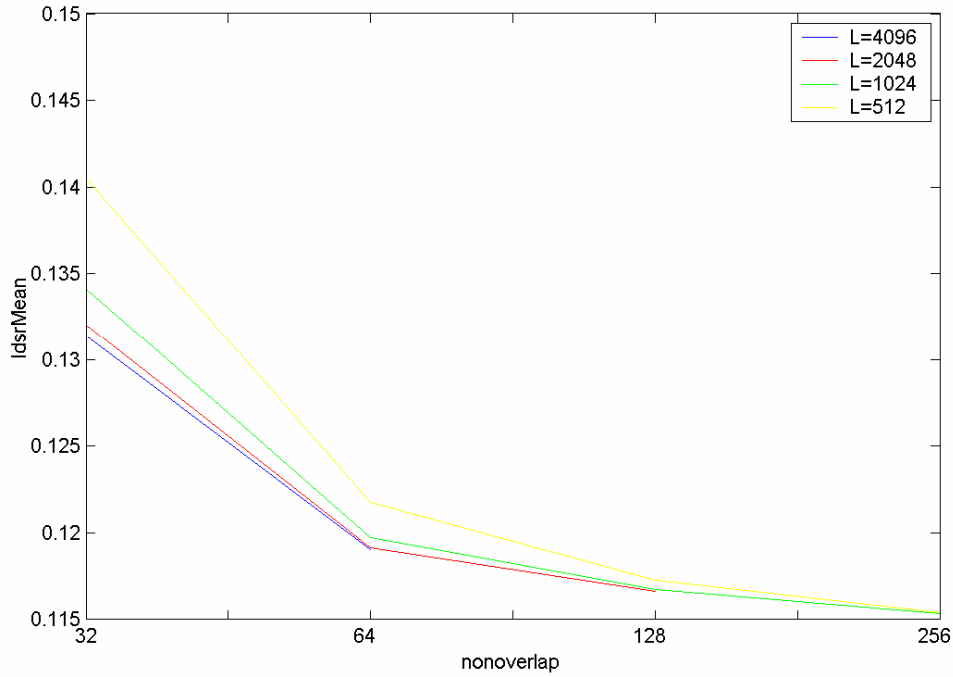
Pro tyto signály jsem spočítal IDSR v závislosti na parametru τ a separovaný signál řeči s potlačenou hudbou $\hat{s}^M(t)$. Nejoptimálnější IDSR jsem určil pomocí funkce fminsearch implementované v Matlabu. Tato funkce vyhledává minimum funkcí. V našem případě jde o závislost IDSR na volbě τ.

Je zřejmé, že nejlepší by bylo nastavení co největšího okénka L a naopak co nejmenšího nepřekryvu M. Z literatury jsem zjistil, že nejideálnější hodnoty L a M by se měly pohybovat někde v okolí hodnot L = 1024 a M = 128. Proto jsem testoval nastavení L rovno 512, 1024, 2048, 4096 a M rovno 32, 64, 128, 256. Po skončení experimentu jsem získal 1000 výsledných hodnot. Z těch bylo nutné nakonec určit průměrnou hodnotu, která bude nejvýhodnější pro největší počet signálů. Zkusil jsem výsledek určit pomocí mediánu a průměru. Pomocí obou hodnot jsem nakonec dospěl k velmi podobným výsledkům, které ukázaly, že teoretický předpoklad nastavení L = 1024 a M = 128 je nejvýhodnější.

Z grafů (Graf 1, Graf 2) je patrné, že při nastavení okénka na velikost vyšší než 1024 už dochází pouze k nepatrným změnám optimálního IDSR, tudíž při použití většího okénka dojde už jen k neznatelnému zvýšení přesnosti výpočtu. Naopak čím větší je velikost okénka, tím se výrazně zvyšují nároky na výpočtový čas. Proto se volba okénka L = 1024 jeví jako nejefektivnější.

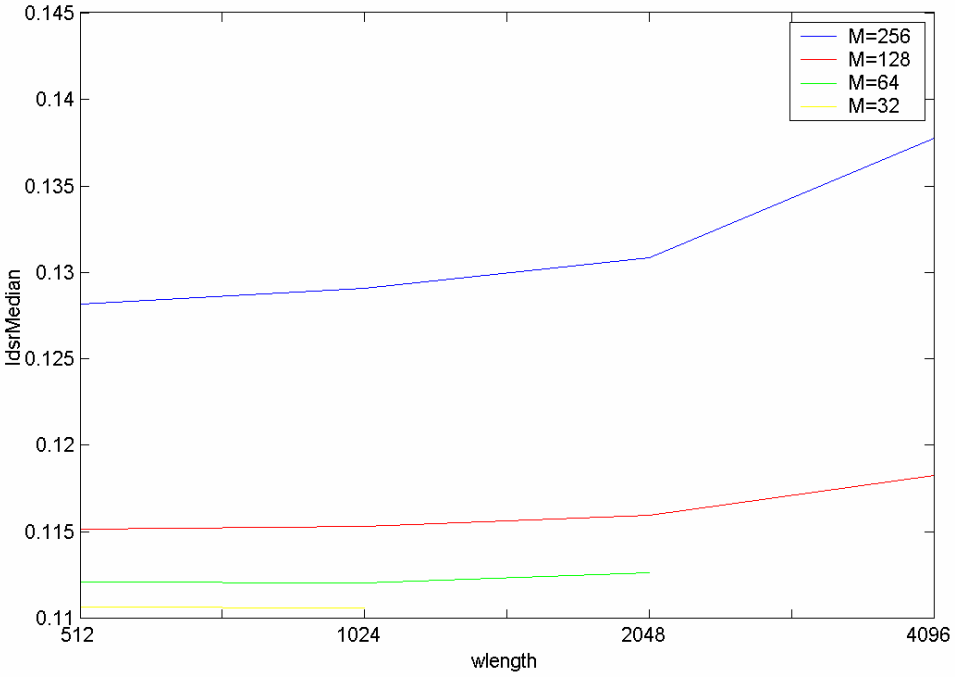


Graf 1: Závislost mediánu IDSR na nepřekryvu okének

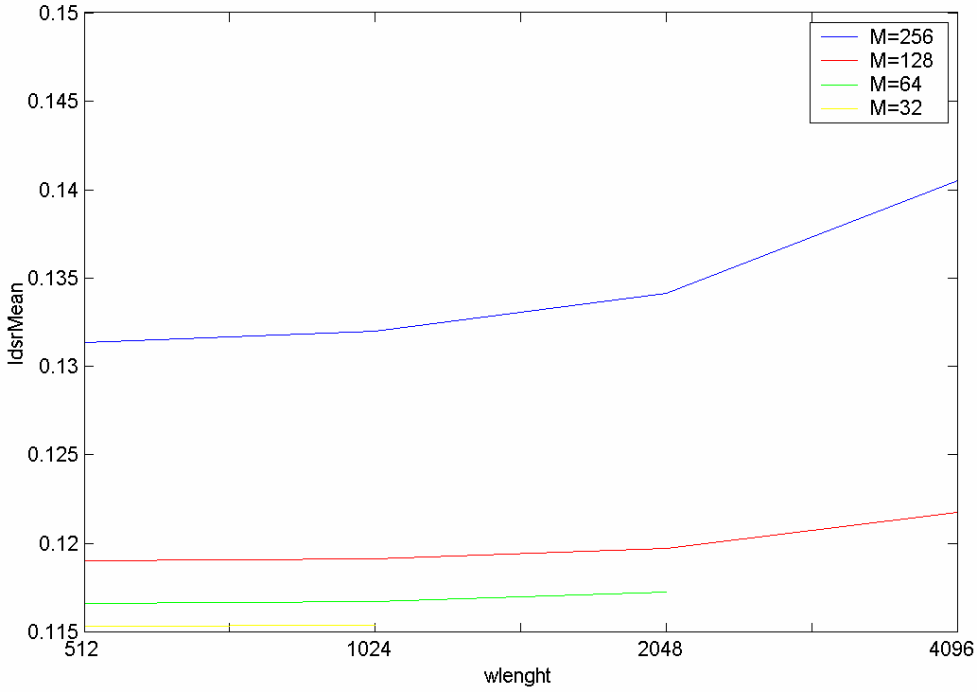


Graf 2: Závislost průměrné hodnoty IDSR na nepřekryvu okének

Jak už je výše uvedeno, máme snahu z důvodů přesnosti nepřekryv volit co nejnižší. Z níže uvedených grafů (Graf 3, Graf 4) můžeme vidět, že do velikosti nepřekryvu M = 128 se IDSR výrazně zlepšuje. Dalším snižováním nepřekryvu je už zlepšení zanedbatelné a obdobně jako u volby okénka se výpočet stává výrazně časově náročnějším. S ohledem na efektivnost výpočtu se tedy volba M = 128 ukazuje jako nejlepší. V grafech (Graf 3, Graf 4) nejsou zaneseny hodnoty pro M = 128 při velikosti okénka L větší než 2048 a pro M = 32 pro velikosti okénka větší než 1024 proto, že z důvodů výpočtové náročnosti nebylo možné určit IDSR numericky.



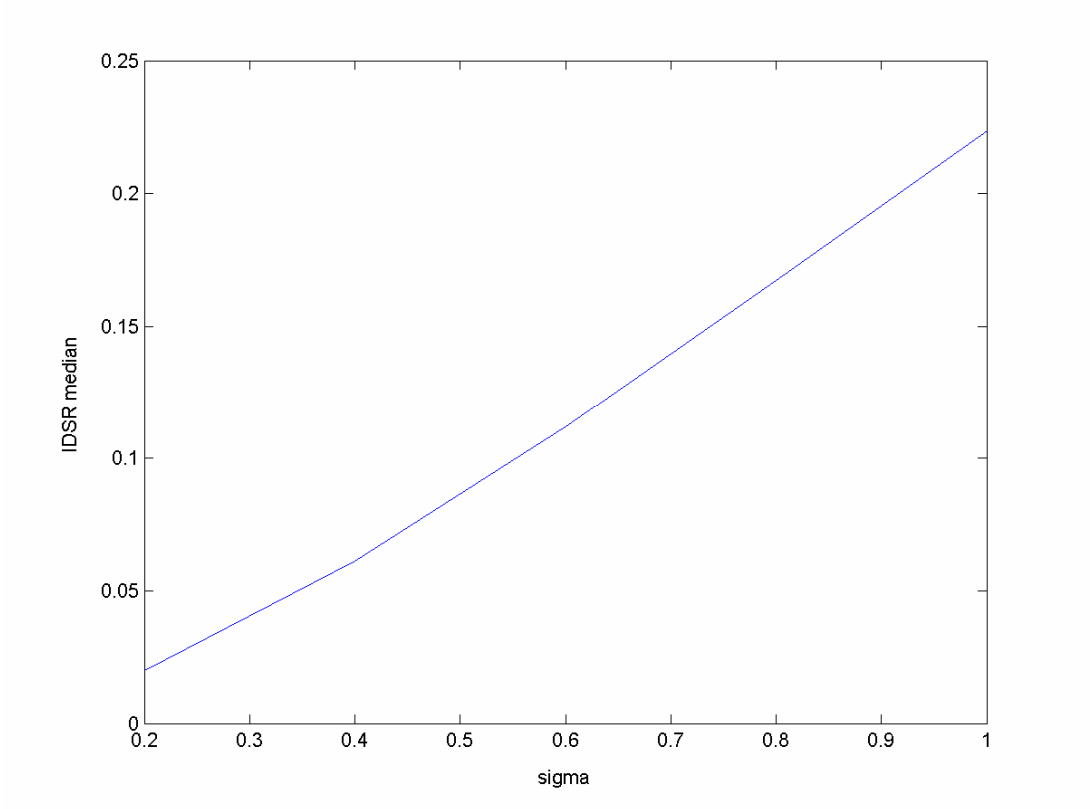
Graf 3: Závislost mediánu IDSR na velikosti okénka



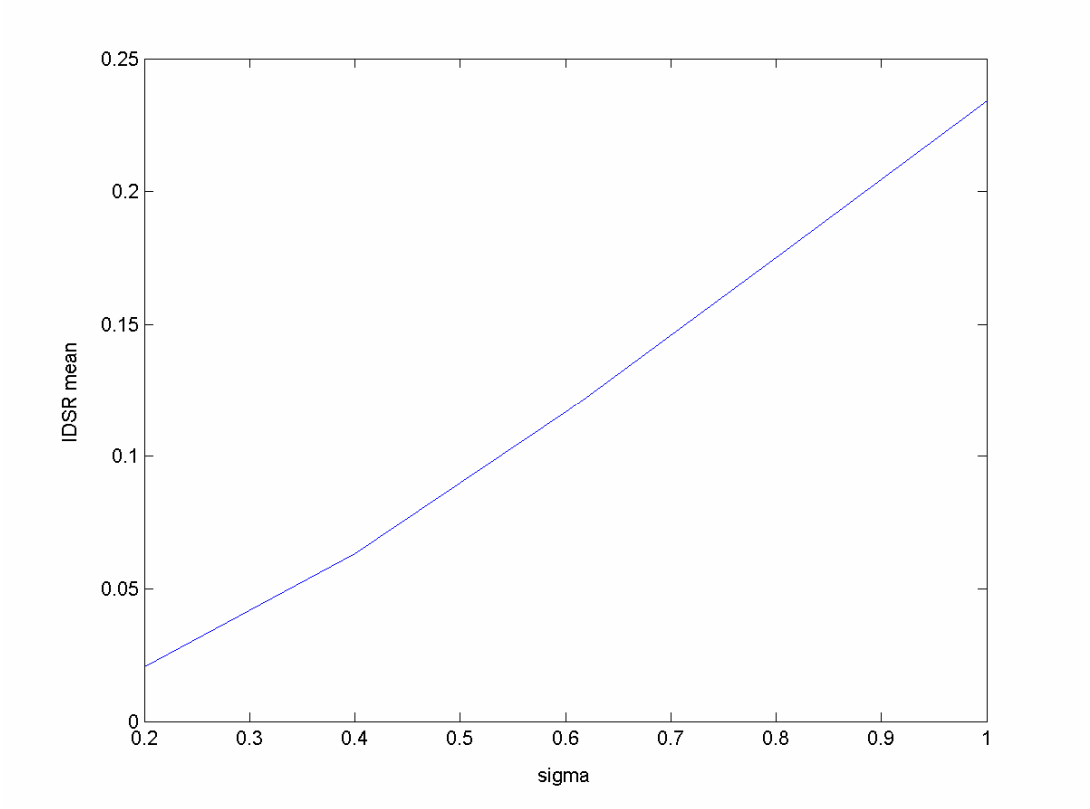
Graf 4: Závislost průměrné hodnoty IDSR na velikosti okénka

Současně jsem testoval, jak bude výsledek ovlivněn intenzitou zarušení σ. Zkoušel jsem provádět výpočet pro různé hodnoty sigma od 0,2 po dvou desetinách až do 1. Výsledné hodnoty IDSR rostou, ale zůstává zhruba zachován jejich vzájemný poměr. Z toho vyplývá, že σ nemá na určení L a M téměř žádný vliv. Výše uvedené grafy (Graf 1 – Graf 4) odpovídají volbě σ = 0,8.

Pro úplnost zde ještě uvedeme závislost optimálního IDSR na σ pro zvolené nastavení L = 1024 a M = 128. Z grafů (Graf 5, Graf 6) je patrné, že optimální IDSR roste se σ téměř lineárně. Grafy ukazují průměrné hodnoty z 1000 provedených pokusů.



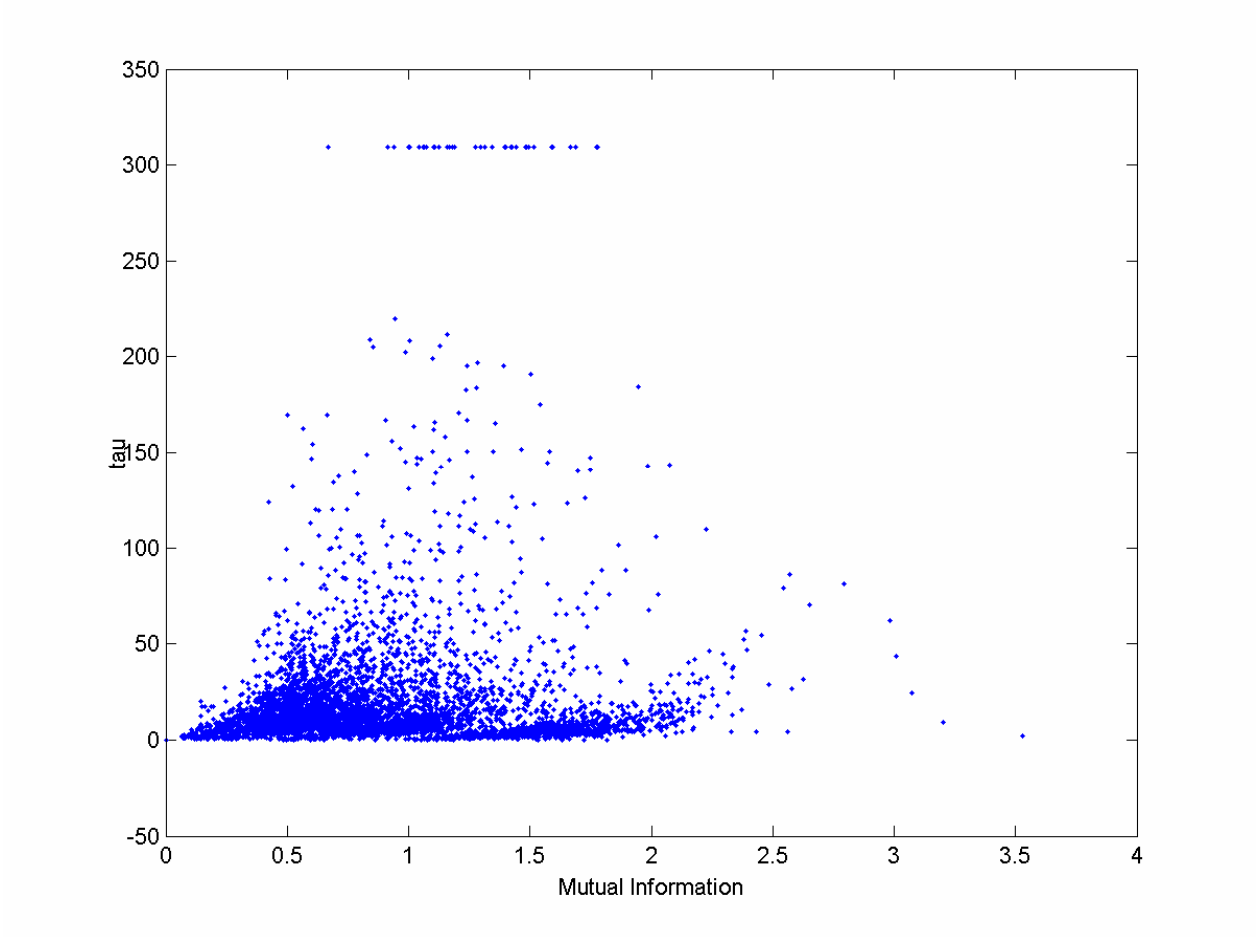
Graf 5: Závislost mediánu IDSR na parametru σ pro nastavení L = 1024, M = 128



Graf 6: Závislost průměrné hodnoty IDSR na parametru σ pro nastavení L = 1024, M=128

4.3. Adaptivní volba prahového parametru τ

Při experimentu prováděném pro určení nejvýhodnější velikosti okének a jejich překryvu, jsem pro každý výpočet ukládal hodnoty optimálního τ^{opt} odpovídajícího optimálnímu IDSR. Současně jsem počítal i vzájemnou informaci signálů $x_1(t)$ a $x_2(t)$. Tyto výsledky jsem využil i při dalším experimentu, jehož cílem byla optimální volba τ^{opt}. Vzájemnou informaci a τ^{opt} jsem určil pro každý signál i pro každé σ. Na následujícím obrázku (Graf 7) vidíme graf rozložení závislosti τ na vzájemné informaci.



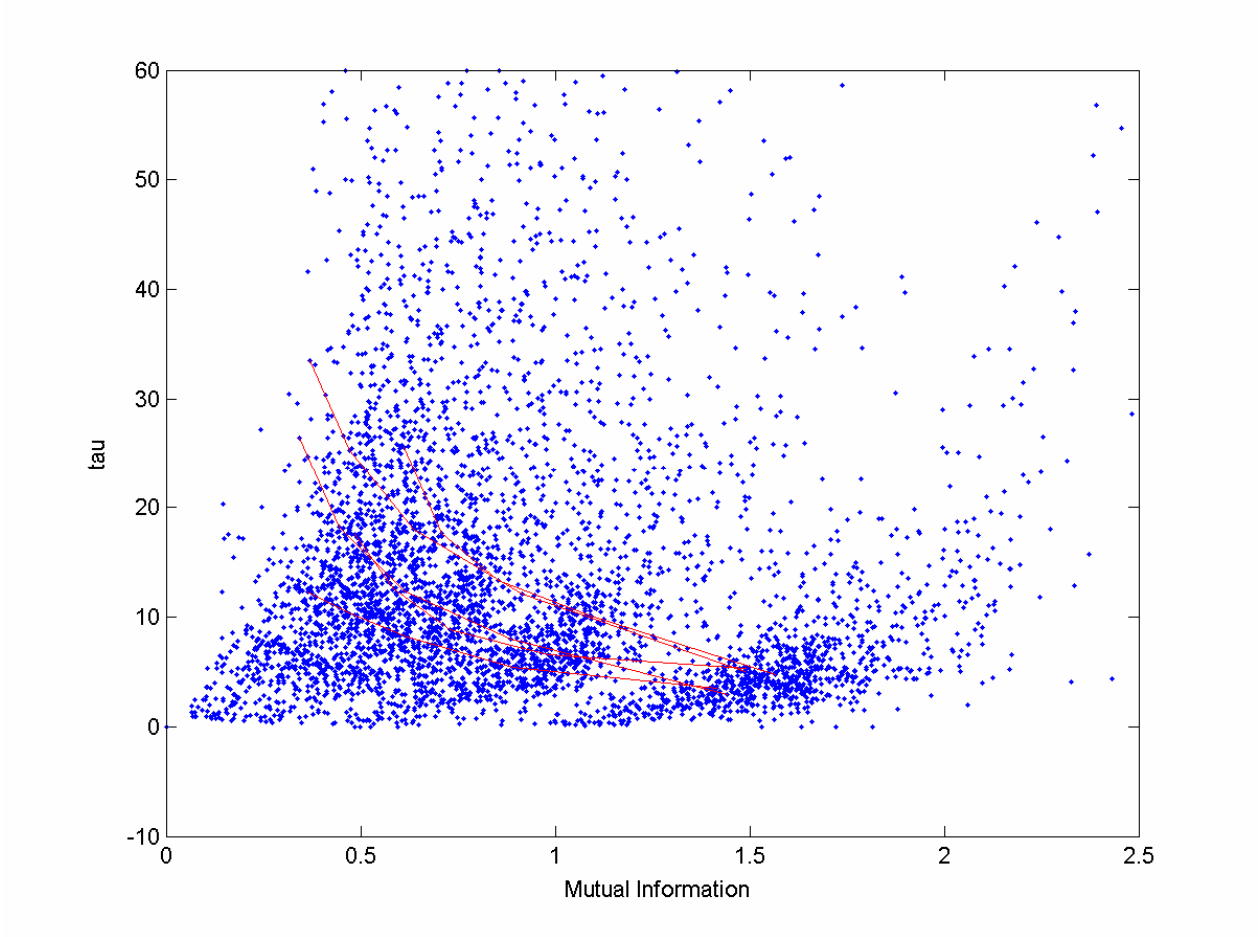
Graf 7: Rozložení τ v závislosti na vzájemné informaci I(x)

Po oříznutí nevyhovujících hodnot jsem získal graf rozložení závislosti τ na vzájemné informaci. Tuto závislost se snažíme popsat, jak již bylo výše uvedeno nějakou funkční závislostí (31):

$$\tau = f(I(x))$$

kde I(x) je vzájemná informace. Jak je vidět z grafu (Graf 7), hodnoty mají značný rozptyl, proto určení funkčního vztahu určujícího τ v závislosti na vzájemné informaci, která by vyhovovala maximálnímu množství signálů, je dost obtížné. Z dalšího grafu (Graf 8) můžeme vidět, že zanedbáním výrazně odlišných hodnot sice zamezíme zkreslení výsledku, ovšem k usnadnění určení vztahu popisujícího závislost parametru τ na vzájemné informaci nedojde. Červeně je v grafu pro názornost zvýrazněno několik náhodně zvolených závislostí τ na vzájemné informaci. Z nich je jasně vidět, že τ s rostoucí vzájemnou informací klesá. τ je na vzájemné informaci nepřímo závislé.

Když se vzájemná informace vstupních signálů $x_1(t)$ a $x_2(t)$ blíží nule, jsou signály maximálně nezávislé. To znamená, že obsahují pouze minimum signálu s(t). Z toho vyplývá, že tuto část signálu x(t) chceme potlačit, a proto potřebujeme vysokou hodnotu prahového parametru τ. Naopak, když je vzájemná informace vstupních signálů $x_1(t)$ a $x_2(t)$ vysoká, znamená to, že energie signálu s(t) obsaženého v x(t) je také vysoká a tuto část signálu chceme zachovat nebo potlačit jen minimálně. Proto je zapotřebí nízká hodnota τ.



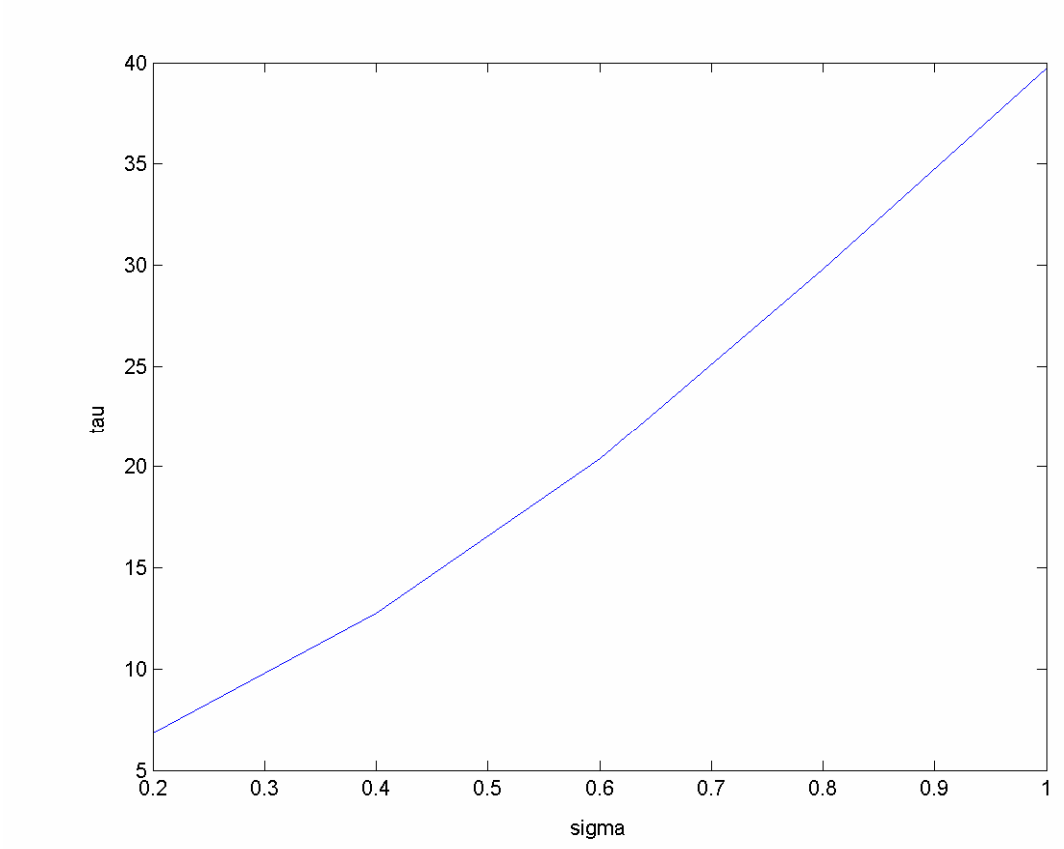
Graf 8: Rozložení τ v závislosti na vzájemné informaci I(x), po odstranění zavádějících hodnot

4.3.1. Určení funkční závislosti prahového parametru τ

Vzájemná informace vstupních signálů $I(x_1, x_2)$ je největší při minimální intenzitě zarušení, to znamená při nízkém parametru σ. To je způsobeno tím, že signál s(t) je v obou signálech $x_1(t)$ a $x_2(t)$ stejný a má výrazně větší intenzitu než y(t), a proto jsou $x_1(t)$ a $x_2(t)$ značně závislé. S rostoucím σ vzájemná informace klesá a stoupá vzájemná nezávislost signálů $x_1(t)$ a $x_2(t)$. Z toho vyplývá, že je vzájemná informace zřejmě závislá na převrácené hodnotě parametru σ. Tuto závislost se snažíme popsat vztahem (34):

$$I(x_1, x_2) = k \cdot \frac{1}{\sigma} \tag{34}$$

Intenzitu zarušení ani míru nezávislosti vstupních signálů předem neznáme, jsme ale schopni určit jejich vzájemnou informaci. Proto se dále nabízí myšlenka určit prahový parametr τ nějakým způsobem právě ze vzájemné informace vstupních signálů. Tím dosáhneme toho, že volba parametru τ pro jednotlivé signály se bude blížit optimální hodnotě τ^{opt}. V následujícím grafu (Graf 9) vidíme, jak závisí průměrná hodnota τ^{opt} na velikosti zarušení σ. Velikost optimálního τ^{opt} roste v závislosti na σ téměř lineárně. Čím více je signál zarušený, tím potřebujeme při maskování potlačit více signálu, a proto potřebujeme použít vyšší hodnotu τ.

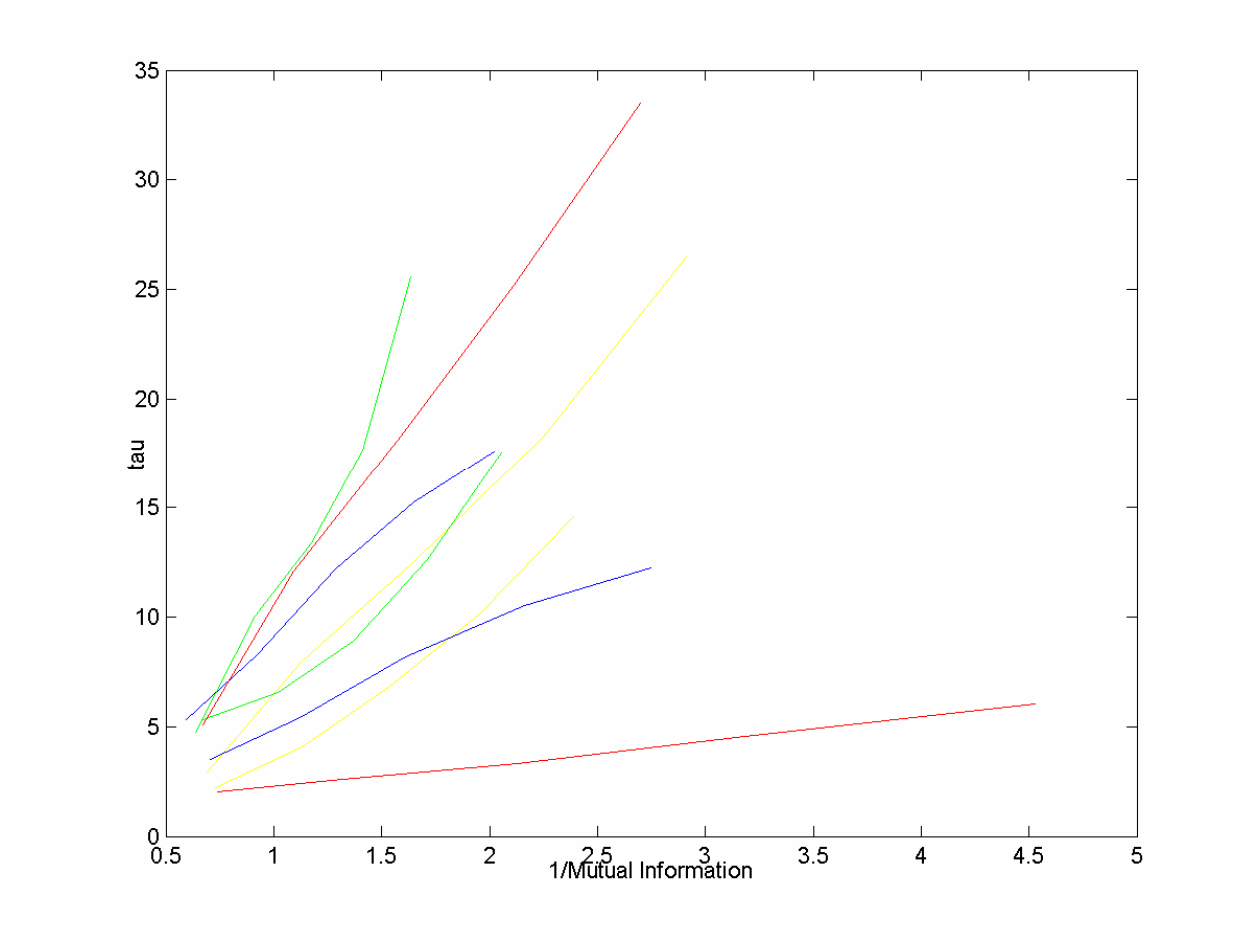


Graf 9: Závislost τ na parametru σ

Budeme uvažovat závislost parametru τ na vzájemné informaci. Experiment ukázal (viz. Graf 10), že závislost τ na vzájemné informaci nejlépe odpovídá relativně jednoduchému funkčnímu vztahu y = 1/x násobeného nějakou konstantou, a to s dostatečnou přesností. Vztah tedy bude vypadat nějak takto:

$$\tau = a \cdot \frac{1}{I(x)} \tag{35}$$

kde I(x) je vzájemná informace signálů $x_1(t)$ a $x_2(t)$ a *a* je experimentálně zjištěný parametr. Naším úkolem je nyní určit tento parametr tak, aby vztah odpovídal maximálnímu množství signálů. To je značný problém, protože signály jsou velmi odlišné, a tím pádem se značně liší i jednotlivé závislosti parametru τ na vzájemné informaci.



Graf 10: Závislost τ na převrácené hodnotě vzájemné informace pro několik náhodně zvolených signálů

Na obrázku (Graf 10) je pro názornost ukázáno osm náhodně zvolených závislostí τ a x(t) na převrácené hodnotě vzájemné informace 1/I(x). Můžeme vidět, že takto chápaná závislost má pro všechny měřené signály zhruba lineární průběh, výrazně se lišící pouze směrnicí. Z toho je zřejmé, že aproximace funkční závislosti lineární funkcí je nejvýhodnější. Je však nutné vhodně zvolit směrnici *a*.

Vhodnou směrnici jsem určil tak, že jsem spočítal směrnici pro každý z 1000 signálů vytvořených pro předchozí experiment. Z takto získaných výsledků bylo nutné

statisticky určit nějaký vhodný kompromis, protože výsledky byly samozřejmě podle předpokladu značně odlišné. Pro toto rozhodnutí se nabízela dvě kritéria a to průměr, nebo medián. Pro ověření jsem určil výslednou směrnici pomocí obou. Výsledky obou kritérií jsou velmi blízké, z čehož jsem usoudil, že použití jedné z těchto hodnot pro další práci je vhodné.

Výpočtem mediánu jsem získal: $a = 5.9944$

Výpočtem průměrné hodnoty: $a = 5.8939$

Konečný vztah pro adaptivní volbu prahového parametru τ tedy bude vypadat následovně:

$$\tau = 5{,}8939 \cdot \frac{1}{I(x)} \tag{36}$$

4.3.2. Souvislost τ^{opt} s volbou okénka a překryvu

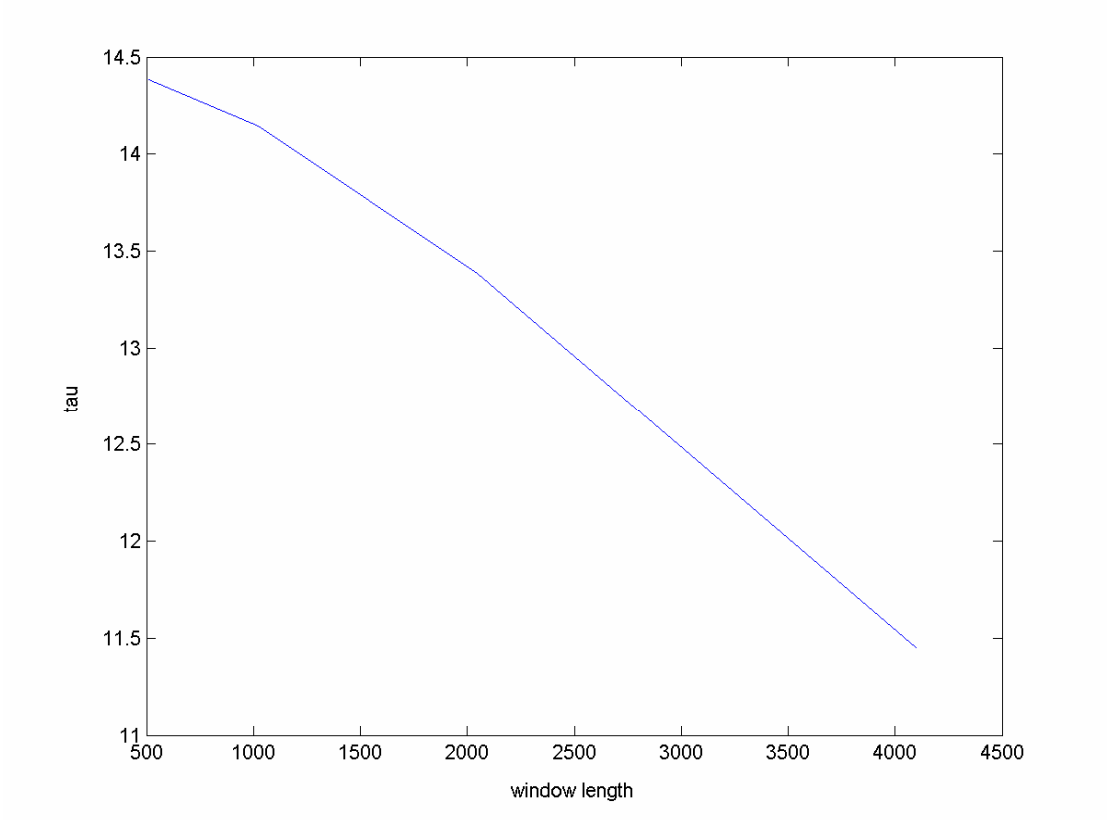
Jak již bylo uvedeno, separace by měla být nejkvalitnější s použitím co největšího okénka a naopak minimálního nepřekrytí. Jak vidíme z grafů (Graf 11, Graf 12), maximální velikost okénka a minimální nepřekrytí vyžaduje, aby nejideálnější hodnota τ^{opt} byla nízká. V grafech (Graf 11, Graf 12) jsou znázorněny průměrné závislosti pro hodnoty parametru $\sigma = 0{,}8$.

V Grafu 11 je zobrazena průměrná hodnota τ^{opt} a její růst se zmenšujícím se okénkem. S rostoucí velikostí okénka určujeme vzájemnou informaci pro větší úsek signálu, a proto jsme schopni ji určit přesněji a parametr τ je potřeba nastavit na nižší hodnotu.

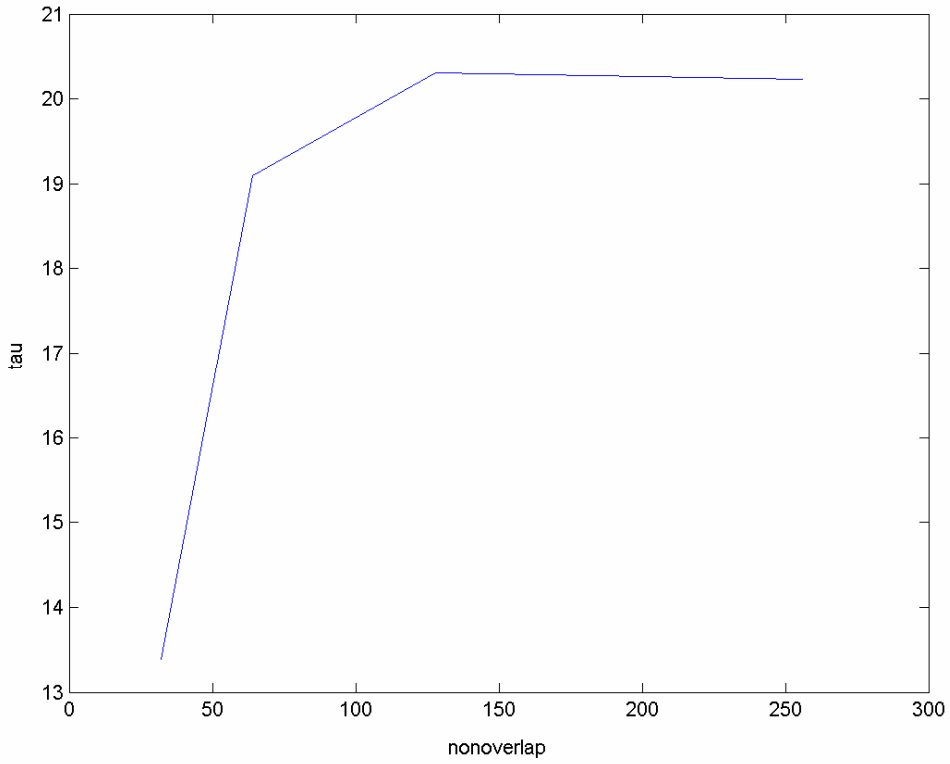
V Grafu 12 je znázorněno, jak průměrná hodnota ideálního nastavení τ^{opt} naopak roste s velikostí nepřekrytí, a to především v nízkých hodnotách. To je způsobeno tím, že s rostoucím nepřekryvem okének naopak vzájemná informace úseků signálů $x_1(t)$ a $x_2(t)$ klesá, a proto je zapotřebí většího τ.

Tento jev můžeme vysvětlit takto: Čím máme větší okénko, tím větší je i nezávislost segmentu signálů v okénku obsaženém, protože v delším signálu se objeví

více odlišností. Z toho plyne, že vzájemná informace klesá, a proto, z důvodů uvedených již v kapitole 4.4.1., je nutná volba vyššího prahového parametru τ. V případě nepřekrytí je situace obdobná. Čím je velikost nepřekryvu menší, tím je snazší separace signálů a je potřeba volit menší i τ.



Graf 11: Závislost τ^{opt} na velikosti okénka



Graf 12: Závislost τ^{opt} na nepřekryvu okének

5. Srovnání kvality navržené metody

Pro ověření navržené metody separace řeči jsem provedl experiment, pomocí něhož jsem ověřil, nakolik je metoda účinná a dokáže potlačit hudbu v pozadí záznamu, bez poškození původního signálu s(t).

5.1. Příprava experimentu

Pro experiment jsem získal databázi monofonních záznamů různých, převážně zpravodajských pořadů čtyř českých celoplošných televizí. Konkrétně ČT1, ČT2, Nova a Prima. Databáze obsahovala 653 záznamů o různých délkách, od několika sekund až po záznamy v délkách jednotek minut. Jednalo se o databázi obsahující směs záznamů různých řečníků, mužů i žen. Součástí databáze byly také textové soubory odpovídající jednotlivým záznamům, které obsahovaly jejich ručně vytvořený přepis.

5.2. Vlastní experiment

Pro provedení experimentu bylo zapotřebí ke každému záznamu v databázi vytvořit zarušený záznam obdobným způsobem jako při návrhu parametrů metody. K zarušení záznamů byla použita stejná databáze hudebních záznamů ve formátu .wav. Z těchto zarušených záznamů jsem pomocí navržené metody získal zpět původní signál s(t) respektive separovaný signál $\hat{s}^M(t)$. Prakticky se jednalo o návrh reálného skriptu pro separaci skutečných záznamů.

1) Mixování signálů

Mixování jsem prováděl obdobným způsobem jako při návrhu parametrů, kdy jsem taktéž potřeboval vytvořit uměle zarušené signály, u nichž jsem předem měl informaci o původním signálu s(t). Jediná odlišnost byla v tom, že jsem nevytvářel pouze vzorky o délce 2^{15}, ale mixoval jsem celé záznamy. Všechny záznamy řeči v databázi byly kratší než hudba, kterou jsem přičítal. V případě, že tomu tak u

některého záznamu nebylo, hudební záznam jsem ve smyčce zopakoval. Tím bylo zajištěno, že vždy bude hudební záznam delší než řeč a tudíž dojde k zarušení celého signálu s(t).

2) Normování délky signálů

Signál s(t) bylo nutné doplnit nulami tak, aby se jeho délka shodovala s délkou hudby. Následně jsem celý signál x(t) ořízl tak, aby byl záznam co nejkratší a aby délka byla rovna nějakému násobku 2^{15}. Zároveň aby zůstala zachována celá délka signálu s(t) a nedošlo ke ztrátě jeho konce. Zkrácení bylo potřebné proto, abychom se z konce signálu x(t) zbytečně nesnažili získat s(t), které už tam není obsaženo. Tím pádem je pro nás bezcenný a pouze zbytečně zvyšuje výpočtový čas algoritmu separace.

3) Vytvoření signálů u(t) a v(t)

Po provedení bodů 1) a 2) jsem dostal signály x(t), které bych měl v reálné aplikaci jako vstupní, a mohl jsem tedy přistoupit k vlastní separaci. Nejprve jsem si vytvořil signály u(t) a v(t) podle výše zmiňovaných pravidel.

4) Maskování

Separaci jsem prováděl po úsecích o délce 2^{15}. Signály jsem rozdělil na několik částí o této délce. Proto bylo vhodné oříznout x(t) na délku rovnou násobku 2^{15}, aby i poslední část měla přesně tuto délku. Separaci jsem provedl pro každou část zvlášť a poté jsem úseky zase spojil ve správném pořadí dohromady. Pro každou část jsem nejprve určil vzájemnou informaci, z níž jsem podle výše uvedeného vztahu (36) určil τ. Poté jsem maskováním pomocí spojité masky (20) získal separovaný signál $\hat{s}^M(t)$.

5) Zápis signálů do souborů

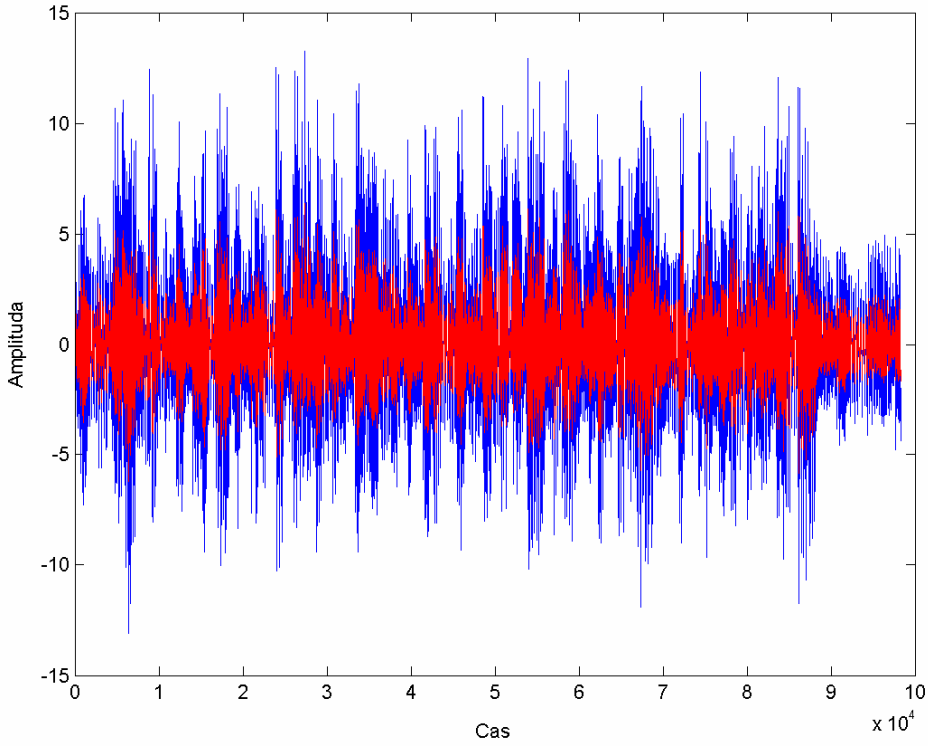
Na závěr jsem z jednotlivých úseků opět poskládal signál $\hat{s}^M(t)$ a zapsal ho do databáze pod stejným názvem souboru, ale do jiné složky kvůli možnosti identifikace a srovnání signálů. K tomu jsem ještě opět pod stejným názvem ukládal monofonní signál vzniklý součtem $x_1(t)$ a $x_2(t)$, což reprezentovalo zarušený signál.

6) Ověření úspěšnosti separace

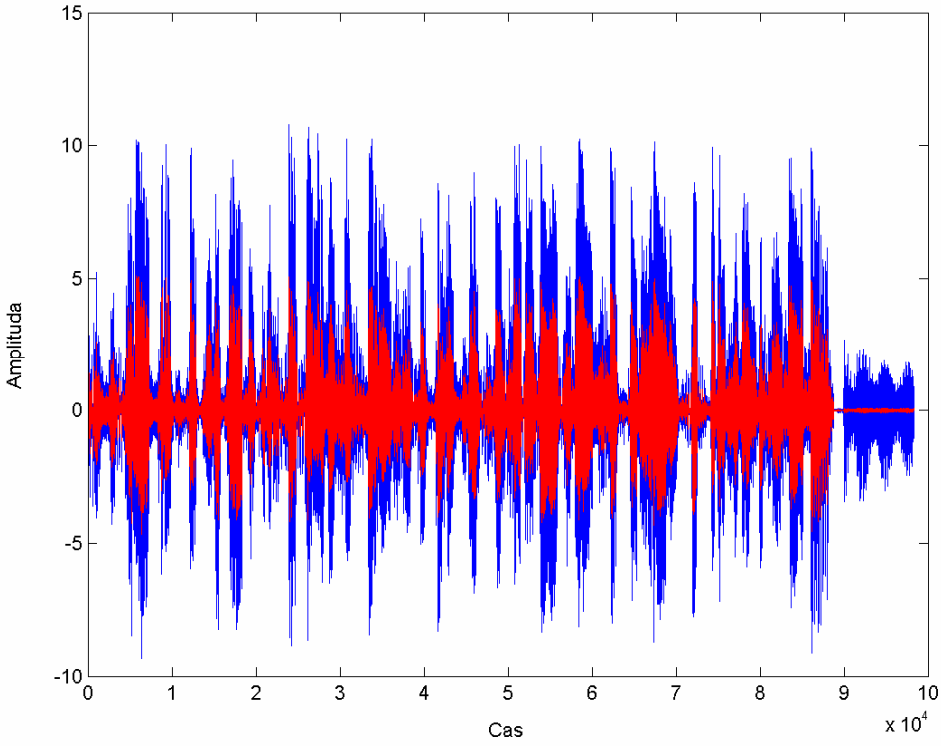
Takto vytvořenou databázi, společně s textovými soubory obsahující přepis řeči, jsme nechali vyhodnotit automatickým rozpoznávačem řeči. Ten je schopen v případě, že mu dodáme text, který je v záznamu obsažen, určit i procentuelní úspěšnost rozpoznání řeči. Předpokládáme nejvyšší procento úspěšnosti u původního signálu s(t). Úspěšnost jeho rozpoznání by se měla pohybovat někde kolem 80%. Úspěšnost rozpoznání zarušeného signálu x(t) se pravděpodobně bude pohybovat někde v okolí 30%, maximálně 40%. Úspěšnost rozpoznání získaného signálu $\hat{s}^M(t)$ by se měla v průměru pohybovat někde mezi těmito hodnotami. Minimum určitě neklesne pod hodnotu zarušeného signálu, ale nemůže být vyšší než u původního nezarušeného signálu s(t). Naším cílem je samozřejmě se co nejvíce přiblížit úspěšnosti dosažené rozpoznáváním původního nezarušeného záznamu.

5.3. Výsledky experimentu

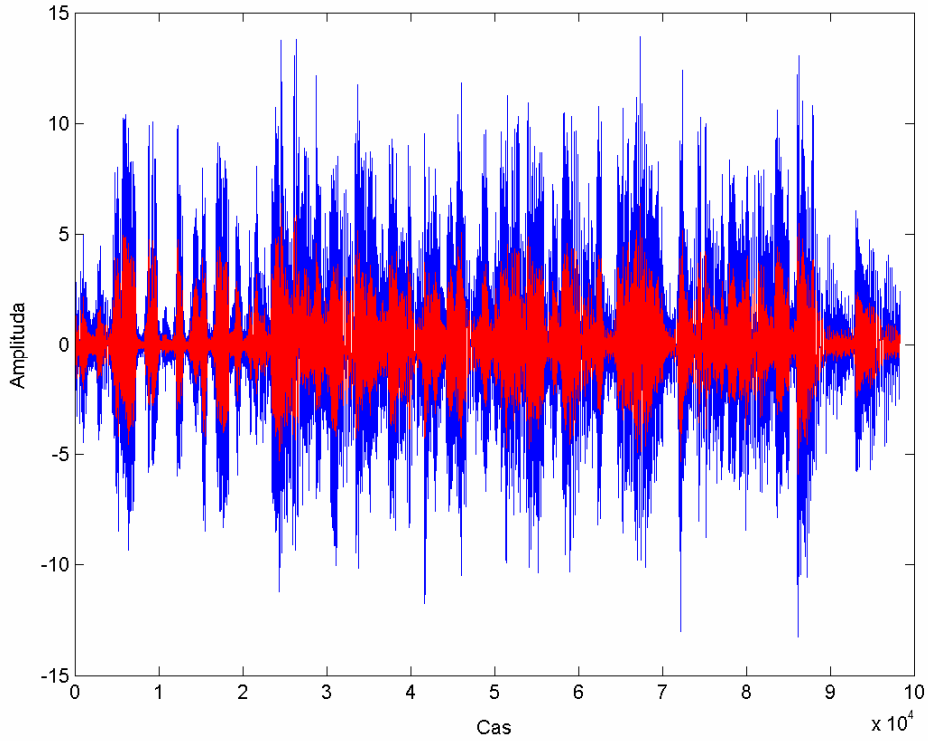
Z ilustrativních obrázků (Obr.3, Obr.4, Obr.5) vidíme srovnání části zarušeného signálu x(t) zobrazeného modrou křivkou a získaného signálu $\hat{s}^M(t)$ zobrazeného červeně. Jedná se o zobrazení průběhu stejného záznamu, za použití různého parametru *a*. Z obrázků je vidět, že k největšímu potlačení hudby dochází při volbě $a = 8$. Na obrázku (Obr.6) je pro porovnání zobrazen samotný nezarušený signál s(t).



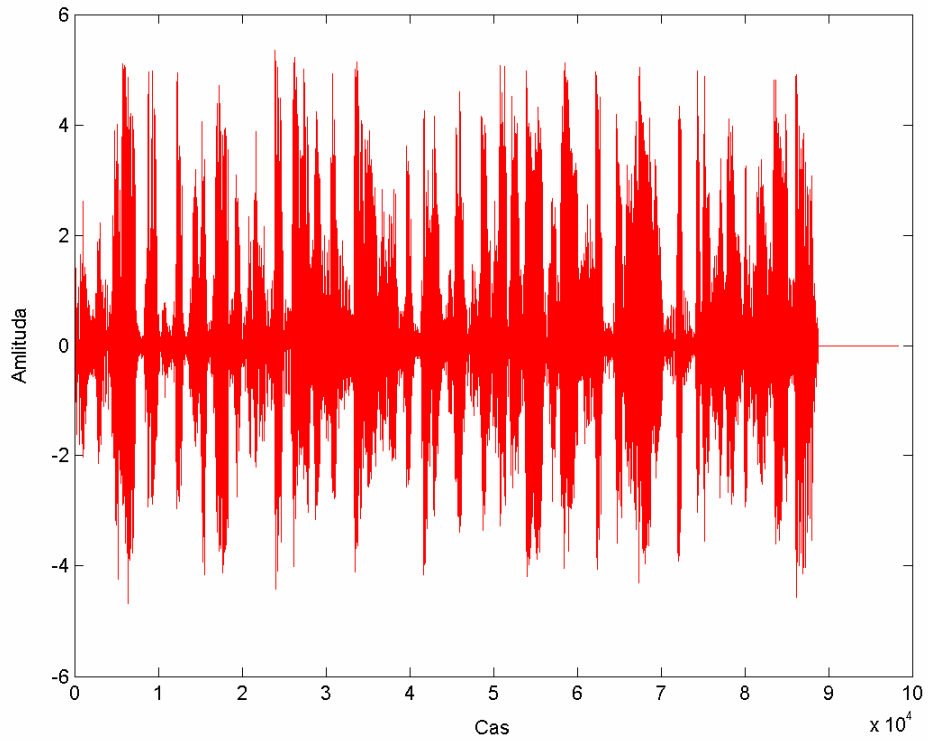
Obr.3: Srovnání x(t) a $\hat{s}(t)$ pro a = 5,89



Obr.4: Srovnání x(t) a $\hat{s}(t)$ pro a = 8

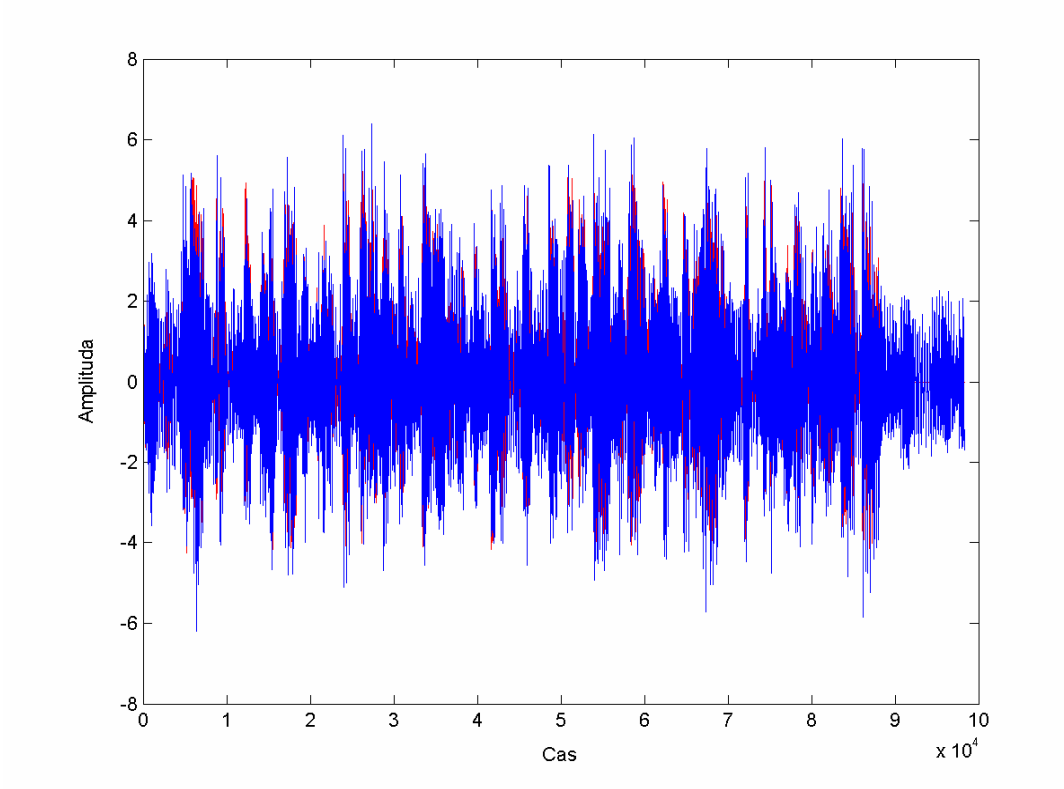


Obr.5: Srovnání x(t) a $\hat{s}(t)$ pro a = 10

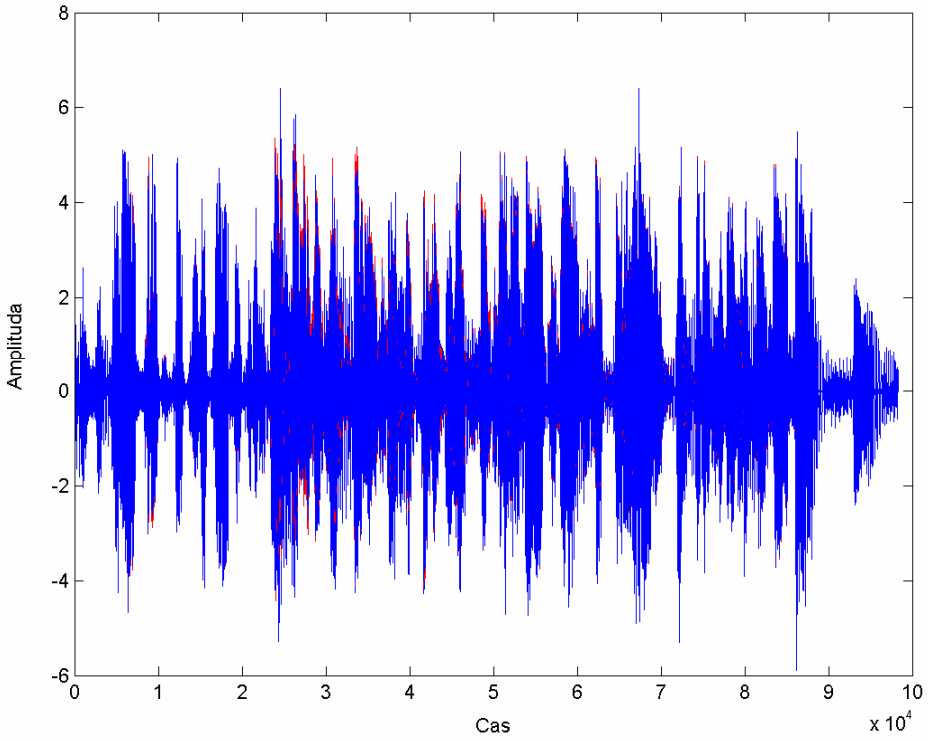


Obr.6: Nezarušený signál s(t)

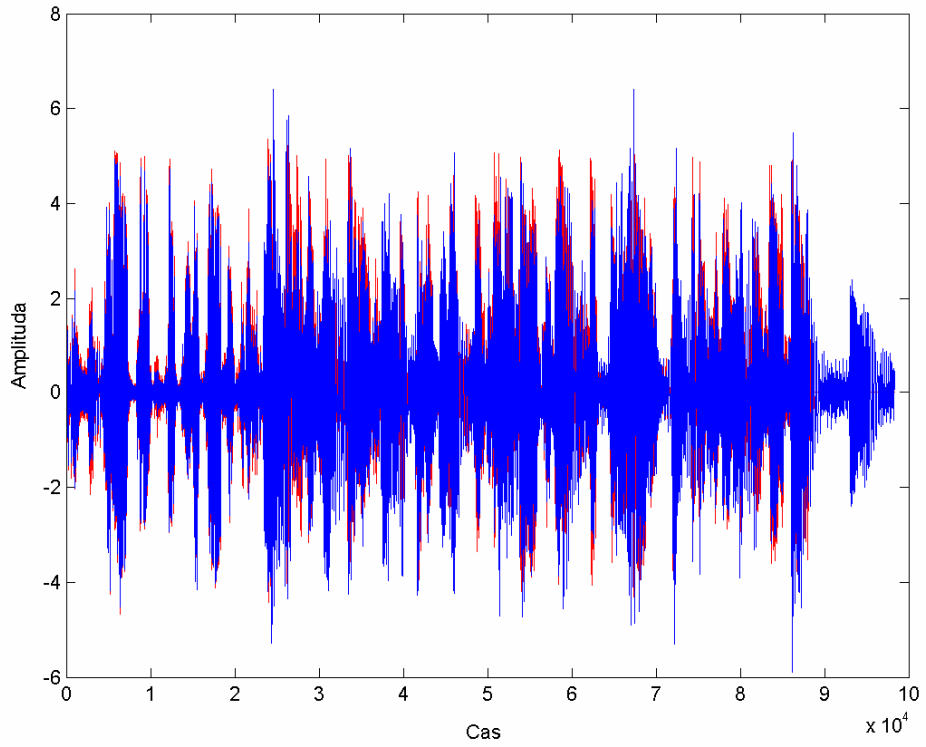
Na dalších třech obrázcích (Obr.7, Obr.8, Obr.9) můžeme vidět ilustrativní srovnání části původního signálu s(t) před zarušením zobrazené červenou křivkou se signálem $\hat{s}^M(t)$ získaným po separaci, který je zobrazen modře. Průběhy jsou téměř totožné, což potvrzuje, že metoda funguje správně podle předpokladů. Nejlépe se průběhy překrývají při nastavení a = 8. To znamená, že zde dochází k nejlepší rekonstrukci původního signálu s(t). Při nastavení a = 5,89 vidíme, že potlačení hudby ještě není úplně dokonalé. Naopak při a = 10 je již získaný signál $\hat{s}^M(t)$ částečně zkreslený.



Obr.7: Srovnání s(t) a $\hat{s}^M(t)$ pro a = 5,89

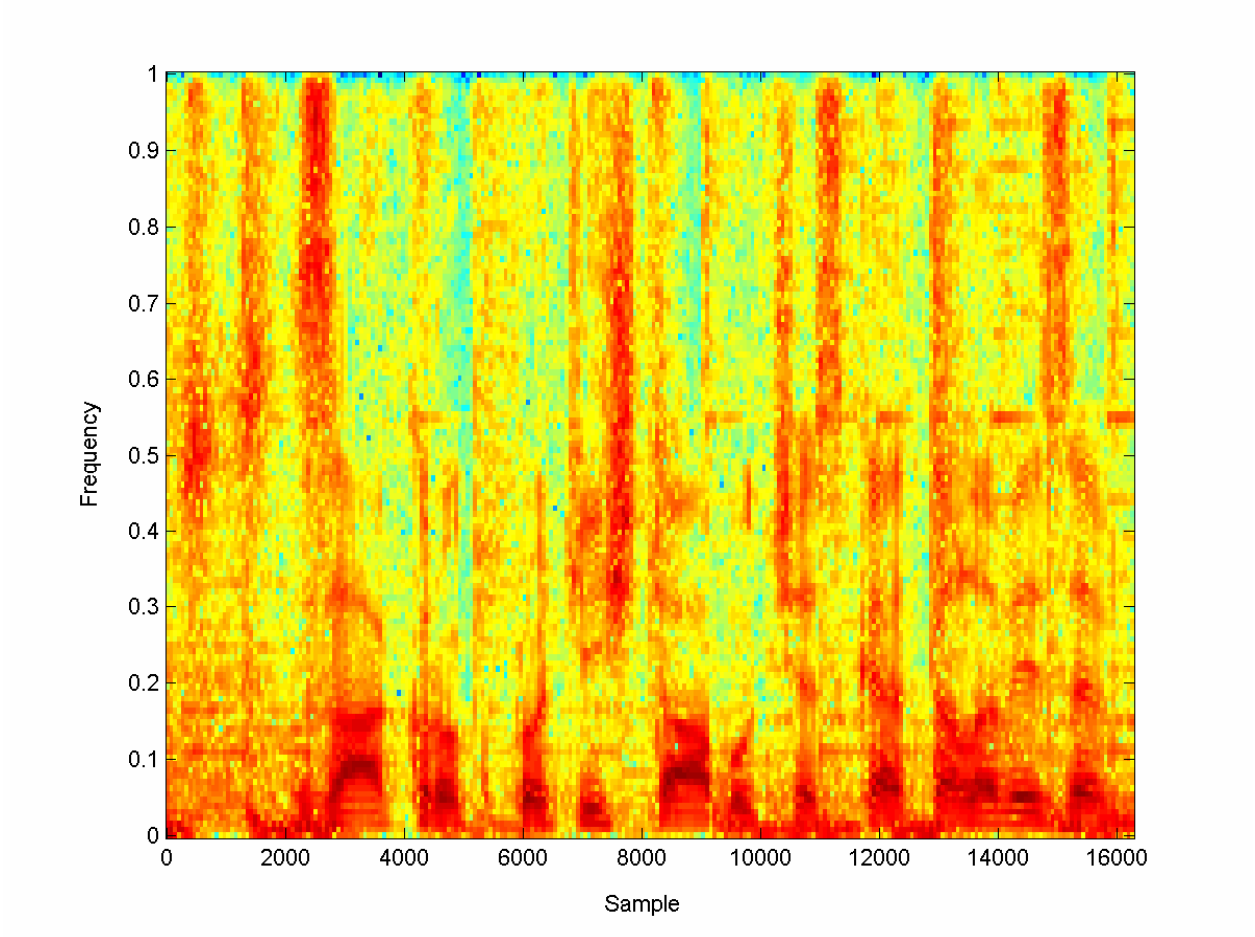


Obr.8: Srovnání s(t) a $\hat{s}^M(t)$ pro $a = 8$

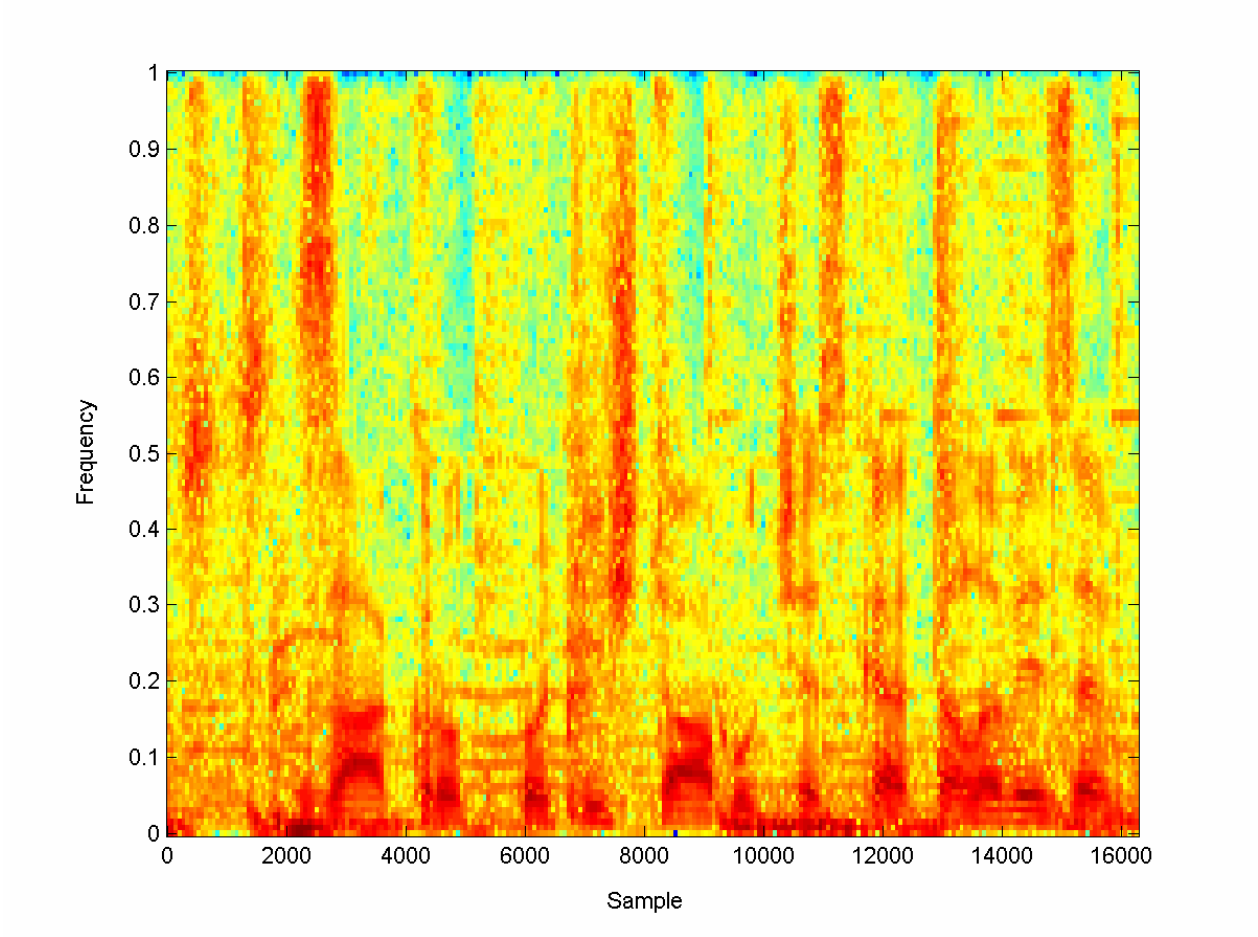


Obr.9: Srovnání s(t) a $\hat{s}^M(t)$ pro $a = 10$

Na následujících třech obrázcích můžeme vidět, jak se změní frekvenční spektrum nezarušeného signálu s(t), zobrazeného na prvním obrázku (Obr.5). Na druhém obrázku (Obr.6) vidíme zarušený signál x(t), respektive součet levé a pravé strany signálu, a na třetím obrázku (Obr.7) vidíme spektrum získaného signálu $\hat{s}^M(t)$ Z porovnání obrázků 6 a 7 je zřejmé potlačení některých kmitočtů, které nejsou v řeči obsaženy. Výrazný rozdíl mezi obrázky 5 a 7 je daný odlišnou intenzitou původního a získaného signálu.

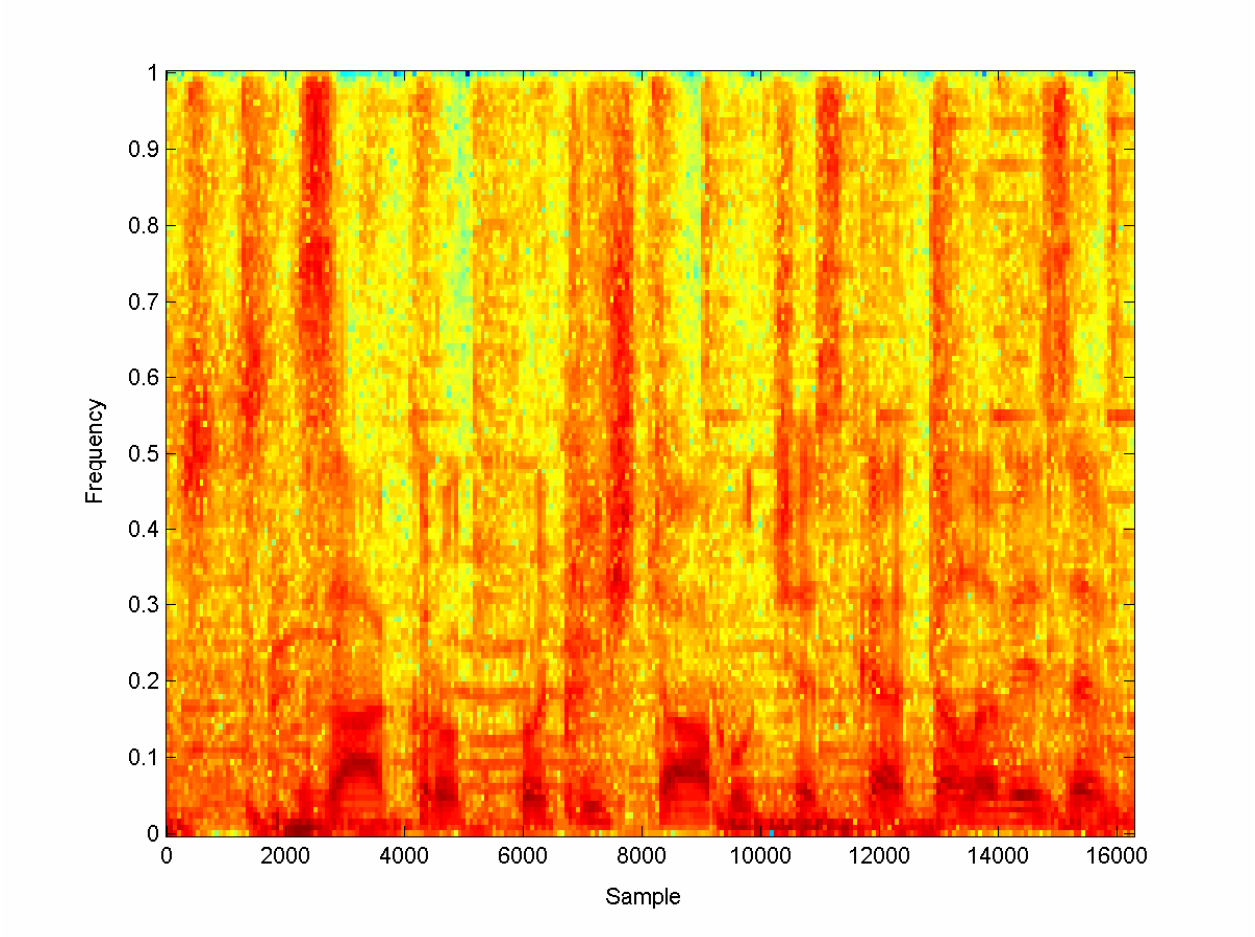


Obr.10: Spektrograf s(t) pro $a = 8$



Obr.11: Spektrograf x(t) pro a = 8

Na dalším obrázku (Obr. 12) vidíme spektrograf signálu $\hat{s}^M(t)$. Na první pohled se frekvenční spektrum jeví výrazně odlišné od spektra původního signálu s(t). To je však způsobeno pouze tím, že jsou zde zastoupeny jiné kmitočty v odlišných poměrech, a proto jsou zobrazeny v jiné barevné škále.



Obr.12: Spektrograf $\hat{s}^T(t)$ pro $a = 8$

5.4. Vyhodnocení experimentu

Experiment nakonec ukázal, že nastavení parametru a ve vztahu určujícím prahový parametr τ (35) na hodnotu přibližně $a = 6$ není nejvhodnější, protože nedojde k dostatečnému potlačení signálu hudby y(t). Čím větší nastavíme a, tím víc roste i prahový parametr τ. Tím pádem se zvyšuje množství signálu který potlačíme.

Zkusil jsem několikrát zopakovat experiment, abych zjistil, jestli je možné a natavit vhodnějším způsobem. Z grafického zobrazení výsledků se ukázalo, že se výsledné signály zlepšují až do hodnoty $a = 8$. Zde je také výsledný signál $\hat{s}^T(t)$ nejlépe srozumitelný pro lidské ucho. Při dalším zvyšování parametru a jsem již získával příliš vysoké τ a tím se zvyšovalo zkreslení získaného signálu řeči $\hat{s}^T(t)$.

Nejvhodnější nastavení tedy bude a = 8. Konečný vztah pro adaptivní volbu prahového parametru τ (36) bude vypadat následovně:

$$\tau = 8 \cdot \frac{1}{I(x)}$$

Tato volba se potvrdila jako správná především podle výsledků z automatického rozpoznávače řeči i vizuálně na zobrazení samotných signálů. Příklad je uveden výše (Obr.3 – Obr.9).

5.4.1. Výsledky automatického rozpoznávání

Z následujících třech tabulek vidíme, že k nejúspěšnějšímu rozpoznávání došlo při volbě a = 8. Při nastavení a = 10 sice došlo k nepatrně většímu zlepšení rozdílu rozpoznávání signálu $\hat{s}^T(t)$ a x(t), ale to je dané tím, že bylo dosaženo horšího rozpoznání zarušeného signálu x(t). Při tomto nastavení již dochází k částečnému potlačení signálu $\hat{s}^T(t)$, proto vyšší hodnota a již nemá smysl. V první tabulce (Tab.1) se oproti dalším dvěma liší úspěšnost rozpoznání zarušeného signálu x(t) a tím i $\hat{s}^T(t)$. To je způsobeno tím, že pro opakování experimentu jsem zvolil jinou hudbu.

Tab.1: Výsledky automatického rozpoznávání při volbě a = 5,89

ID	Průměrná úspěšnost rozpoznávání [%]
Čisté s(t)	80,64
Zarušené x(t)	65,75
Vyčištěné $\hat{s}^T(t)$	67,83

Tab.2: Výsledky automatického rozpoznávání při volbě a = 8

ID	Průměrná úspěšnost rozpoznávání [%]
Čisté s(t)	80,64
Zarušené x(t)	61,9
Vyčištěné $\hat{s}^T(t)$	64,21

Tab.2: Výsledky automatického rozpoznávání při volbě *a* = 10

ID	Průměrná úspěšnost rozpoznávání [%]
Čisté s(t)	80,64
Zarušené x(t)	61,08
Vyčištěné $\hat{s}^T(t)$	63,76

6. Závěr

Závěrem diplomové práce bych chtěl několika slovy stručně shrnout přínos úlohy pro zpracování řeči a možné směry, kterými by se mohla dále vyvíjet. Věřím, že se mi podařilo splnit cíle stanovené v úvodu, a zároveň dostatečně srozumitelně popsat řešený problém.

Algoritmus slepé separace řeči ze stereofonního záznamu, tak jak byl v této diplomové práci navržen, lze dále využít v laboratoři zpracování řeči na Technické univerzitě v Liberci. Byl zde uveden postup pro relativně snadnou a efektivní separaci řeči ze zarušeného záznamu s minimálním zkreslením. Tento algoritmus je možné implementovat jako součást rozsáhlejších projektů zpracovávajících řeč ze záznamů. Kvalita signálu a jeho „vyčištění“ od rušivých signálů nenesoucích žádnou informaci, je jedním ze základů úspěšného počítačového zpracování řeči. Značný význam může mít metoda například pro automatický přepis zvukových nahrávek do textových dokumentů.

Jakým způsobem bychom se mohli ubírat dál? Na problém separace řeči můžeme nahlížet různými úhly pohledu. Řešení úlohy je vždy do určité míry kompromisem, ať už mezi zkreslením řeči a potlačením hudby, nebo mezi rychlostí zpracování a jeho kvalitou. Já jsem se pokusil všechna tato kritéria zvážit a zvolit nejvýhodnější řešení. Můžeme se ale zamyslet nad tím, jak by se výsledky změnily, kdybychom například volili jinou funkční závislost τ, nebo použili stejný vztah a pouze změnili parametr a. Nebo jak by se výsledky změnily při jiné volbě velikosti okének a jejich překrytí.

V případě, že bychom chtěli navrhnout algoritmus pro separaci řeči z monofonní směsi, separaci by bylo nutné provádět na základě jiných předpokladů než je vzájemná informace, protože bychom měli k dispozici pouze jeden signál. To by nás vedlo k zajímavému problému, jakým způsobem volit parametry maskování takovéhoto signálu. Také by bylo výhodné provést takové změny ve stávajícím algoritmu, aby byl použitelný i pro aplikace pracující v reálném čase. Tím bychom získali nástroj, významně se podílející na kvalitě například automatických překladačů nebo aplikací pro hlasové ovládání, a podobně.

Diplomová práce spočívala v návrhu algoritmu a implementaci v Matlabu. Dále by na ni mohla plynule navázat práce, která by spočívala ve vytvoření běžně použitelné aplikace pro separaci řeči z nějakého vstupního záznamu. Například v jazyku C++, nebo Delphi. Takto vytvořená aplikace by již byla snadno použitelná a bylo by možné ji bez větších problémů implementovat do jakéhokoli softwaru.

Literatura

[1] KOLDOVSKÝ, Z., NOUZA, J., KOLORENČ, J.: Continuous Time-Frequency Masking Method for Blind Speech Separation with Adaptive Choice of Threshold Parameter Using ICA. In: International Conference on Spoken Language Processing Interspeech 2006 — ICSLP 2006, September, 2006, Pittsburgh, USA, pp. 2578-2581, ISSN 1990-9772

[2] KOLDOVSKÝ, Z.: Analýza nezávislých komponent v EEG datech, diplomová práce, FJFI ČVUT, 2002.

[3] KOLDOVSKÝ, Z.: Analýza nezávislých komponent, FastICA a přímý výpočet vzájemné informace [online], Rešeršní práce FJFI ČVUT, září 2000, URL: <http://itakura.kes.vslib.cz/zbynek/pubs/reserska.ps>

[4] KOLDOVSKÝ, Z.: Analýza nezávislých komponent - odhad vzájemné informace a separace konvolutorních směsí [online], Rešeršní práce FJFI ČVUT, září 2001, URL: <http://itakura.kes.vslib.cz/zbynek/pubs/vyzkumak.ps>

[5] Domovská stránka Ing. Zbyňka Koldovského, Ph.D. URL: <http://itakura.kes.tul.cz/zbynek/index.htm>

[6] YILMAZ,Ö., RICKARD, S.:Blind Separation of Speech Mixures via Time-Frequency Masking, IEEE Trans. on Signal Processing, vol. 52, no. 7, July 2004

[7] Wikipedia, Hustota pravděpodobnosti[online], URL: <http://cs.wikipedia.org/wiki/Hustota_pravděpodobnosti>